様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２５ 日現在

機関番号：３３９２４
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１１９４２

研究課題名（和文）オントロジー形式による関係アノテーションとその効果の深層学習による検証

研究課題名（英文）Investigation of Ontology-Style Relation Annotation and Its Effects with Deep Learning

研究代表者
佐々木　裕（Sasaki, Yutaka）

豊田工業大学・工学（系）研究科（研究院）・教授

研究者番号：６０３９５０１９

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：　本研究は，文書からの情報抽出を対象に，オントロジー形式のアノテーションにより作成された訓練データの効果を検証するものである．オントロジー形式のアノテーションでは関係用語が対象とする２つの用語をそれぞれdomain，rangeによりリンクする．
主な成果として，交通教則文およびSemEval 2010タスク8データに対して，オントロジー形式のアノテーションを行ったことが挙げられる．作成した独自のデータセットに対して，深層学習モデルによる用語抽出・関係イベント抽出の評価実験を行い，オントロジー形式のアノテーションの効果を明らかにした．本成果はIF=4.3の国際論文誌に掲載された．

研究成果の学術的意義や社会的意義
オントロジー形式の関係アノテーションを考案し，その効果を独自データセットの構築を通して確認した．日本発の新しいアノテーションを提案したことで，本分野の発展に寄与した．また，交通教則文に対して，オントロジー形式アノテーションを適用したデータセットを公開したことも貢献として挙げられる．英語のデータセットに関しても，関係抽出に関する標準データセットであるSemEval 2010 タスク8データの8,000文に対してオントロジー形式のアノテーションを行い，データセットを公開した．オントロジー形式のアノテーションは今後，言語データからオントロジーへの変換の橋渡しとなることが期待される．

研究成果の概要（英文）：　This study examines the effectiveness of training data generated by Ontology-style annotations for extracting information from documents. Ontology-style annotations link two named entities targeted by a relational term with the domain and range links.
The main result is that Ontology-style annotations were conducted on the "Rules of the Road" and SemEval 2010 Task 8 data. Evaluation experiments of named entity extraction and relation/event extraction using deep learning models were carried out to reveal the effectiveness of Ontology-style annotations. The results were published in an international journal with IF=4.3.

研究分野：自然言語処理

キーワード：オントロジー　オントロジー形式アノテーション　固有表現抽出　関係抽出

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

人類の知恵や知識の多くは，文書情報の形で蓄積されており，文書からの知識獲得は人工知能研究の中でも重要な研究テーマのひとつである．従来，文書からの知識獲得技術は，情報抽出や関係抽出技術として取り組まれてきた．情報抽出技術は，1987 年に DARPA により開催された第 1 回 Message Understanding Conference (MUC-1)以来，多くの研究者の注目を集めている．2000 年代以降，バイオメディカル・インフォマティクスの研究の中で，生命科学文献を対象にした関係抽出技術が発展してきた．

関係抽出は，専門用語とその用語間の関係をアノテーション（注釈付け）した文書データ（コーパス）を作成することで，専門家の知識に基づいた知識獲得を以下の 2 つの副問題として解くアプローチをとってきた．

・専門用語の認識問題
・専門用語間の関係抽出問題

文書からの関係抽出のためのアノテーション作業は，特定の分野の特定の対象項目に対して記述したアブストラクトなどの文書を用意し，それらの文書に対して専門家が対象項目に関連した用語を指定し，それらの用語間の関係を正解データとして GUI ツールを用いて構築していく．専門家の高度の知識が必要な場合が多く，大量の正解データの作成は難しい．専門知識に基づく文脈により，用語や関係のアノテーションが決まるため，対象分野の知識をもつ人による作業が必要である．

一方，研究開始当時より，用語抽出および関係抽出の性能は深層学習以前に比べて大きく向上していた．BERT などの大規模文書情報から事前学習された単語や文のベクトル表現を用いることで，用語抽出は 90%を超える性能が報告されていた．2019 年に入って，関係抽出においても，正解の用語抽出結果が入力として与えられており，かつ関係数が数種類程度で正解データが十分にある場合は 80%を超える予測性能が報告されていた．しかし，依然として関係抽出の対象となる関係数が多く，かつ正解データ数が数百～数千程度の場合，関係抽出の性能は 30%～60%程度に留まっていた．これは，関係の抽出には背景知識や離れた単語間に存在する文脈関係の理解が必要な場合が多いためである．特に，自然言語表現は多種多様であり，定型的なパターンではカバーしきれないことが大きな問題となっていた．

２．研究の目的

本研究では，オントロジー形式のアノテーションを実現し，以下の点を明確化することを目的としている．

（１）従来リンクにより関係を記述していたアノテーションを，関係用語の導入により domain と range 等の少数の関係のみのアノテーションに本当に書き換えられるのかを実際のアノテーション作業を通して得る．
（２）従来型のアノテーションとオントロジー型のアノテーションに対して，同じ深層用語モデルを適用したとき，トータルでの関係抽出の性能は向上するのか

関係抽出よりも用語抽出の性能が高いことと，関係の種類が少ないほど関係抽出の性能が高くなることを考慮すると，オントロジー型のアノテーションが性能面では有利であると推測されていたが，このことを実証した研究は存在していなかった．

３．研究の方法

そこで研究では，オントロジー形式の表現に合わせたアノテーション法を考案した．前述の通り，関係抽出は，用語抽出と関係抽出の 2 つの副問題の組み合わせになっており，経験的には用語抽出の性能と関係抽出の性能の積で，入力文～関係抽出までの End-to-End の性能が決まる．

従前の関係アノテーションは，知識表現とのアナロジーで考えると，用語ノードとノード間の関係リンクからなる，意味ネットワークと同じ表現を用いてきた．近年，知識表現に意味ネットワークを使わず，オントロジーが用いられるのは，オントロジーが，(1)Resource Description Framework (RDF)をデータ表現の基礎にしている，(2)ノード間の関係を自由なリンクで結ぶのではなく，関係を Property ノードで表現し，Property が持つ関係は domain と range の 2 つに基本的に制限する，(3)オントロジー上で Description Logic の基づく推論が規定されている，という長所を持つことが挙げられる．　RDF は「<主語>-<述語>-<目的語>」の 3 つ組で情報をモデリングし，RDF に基づく Web Ontology Language (OWL)により，上位・下位関係（subClassOf）や属性関係（domain/range）等が述語として規定されている．

たとえば，「乗用車の定員は 10 人以下」という文に対する従来のアノテーションは図 1 のようになる．図 1 の「乗用車」→CAPACITY→「10 人以下」という関係アノテーションを，本研究では図 2 のように関係用語を中心として関係アノテーションに変更する．

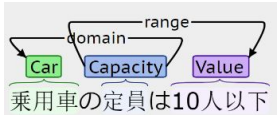

図 1：従来のアノテーション　図 2：オントロジー形式のアノテーション

図 1 のアノテーションはオントロジーの RDF 形式と直接対応し，<乗用車>-<type>-<Car>,<定員>-<type>-<Capacity>，<10 人以下>-<type>-<Value >，　<Capacity>-<domain>-<Car> ，<Capacity>-<range>-<Value> となる．

オントロジー形式でアノテーションすることにより期待される本研究の学術貢献は下記の 3 つである．

（1）オントロジーと同じ形式でアノテーションすることで，関係抽出結果を直接または簡単な変換でオントロジーへの知識追加（population）に利用できる．

（2）従来リンクであった関係に対応する用語を関係用語（オントロジーの Property に対応）としてアノテーションすることで，関係の判断の根拠がアノテーションされる．これにより，従来の関係抽出タスクの担当部分が，高い性能が出ている用語抽出タスクに移管され，かつ関係抽出の対象となる関係の種類が subClassOf，domain，range 等の少数の関係種類になるため，関係抽出の性能も向上させやすい．特に，近年の用語と関係をジョイントで学習する場合，文中の単語数×文中の単語数の行列の学習を関係の種類だけ行う必要があり，従来のように多数の関係を扱うとデータがスパースになり関係抽出の性能が低下するとともに学習時間が大幅に増加する．

（3）オントロジーとアノテーションを直接対応付けることにより，逆にオントロジー中の RDF トリプルを可読性・理解性の高い文の表現で基礎付けることができる．オントロジー中の知識はオントロジーエディターを用いても理解しにくいが，説明文に対するアノテーションをオントロジーの RDF トリプルに対応させておけば，知識と説明文が直接対応し，オントロジーの内容が飛躍的に容易になる．

4．研究成果

主な成果として，交通教則文および SemEval 2010 タスク 8 データに対して，オントロジー形式のアノテーションを行ったことが挙げられる．作成した独自のデータセットに対して，深層学習モデルによる用語抽出・関係イベント抽出の評価実験を行い，オントロジー形式のアノテーションの効果を明らかにした．本成果は IF=4.3 の国際論文誌 Computer Speech & Language に掲載された．

交通教則に対してオントロジー形式のアノテーション適用したデータセットについて，深層学習モデル DyGIE++により固有表現および関係・イベント抽出の評価を行った結果を表 1 に示す．

NER		RE on predicted NEs		RE on gold NEs	
NEClass	F-score	Relation Type	F-score	Relation	F-score
Car	0.499	range	0.58	range	0.616
Road	0.493	location	0.493	location	0.491
Pass	0.489	case	0.454	property	0.479
Driving	0.499	property	0.422	case	0.471
other classes	0.488	other relations	0.419	other relations	0.392
Overall	0.781	Overall	0.482	Overall	0.534

(a) Original RoR corpus

NER		RE on predicted NEs		RE on gold NEs	
NE Class	F-score	Relation Type	F-score	Relation	F-score
Case	0.596	property	0.616	range	0.789
Location	0.596	range	0.605	domain	0.742
Cause	0.591	domain	0.601	property	0.639
Property	0.586	subClassOf	0.446	subClassOf	0.368
other classes	0.586	other relations	0.356	other relations	0.364
Overall	0.747	Overall	0.596	Overall	0.759

(b) New OSR-RoR corpus

NER		RE on predicted NEs		RE on gold NEs	
NE Class	F-score	Relation Type	F-score	Relation	F-score
N/A		location	0.542	location	0.744
		property	0.531	property	0.649
		case	0.448	case	0.560
		cause	0.283	cause	0.470
		other relations	0.203	other relations	0.289
		Overall	0.485	Overall	0.641

(c) After converting predicted OSRs to the conventional RoR relations for the comparison purpose

表 1　交通教則に関する評価結果

表 1(a)は通常のアノテーションの結果である．表 1(b)はオントロジー形式の結果である．これらを比較すると関係抽出(RE on gold NEs)の性能が大きく向上していることがわかる．また，予測された固有表現に対する関係抽出（RE on predicted NEs）についても性能が向上しており，オントロジー形式のアノテーションの効果が明らかになった．表 1(c)はオントロジー形式のアノテーションを従来のアノテーションと同様に domain，range のペアで評価した結果である．こちらも僅かに予測された固有表現に対する関係抽出の性能が向上している．

NER		RE	
Class	F-score	Relation	F-score
e1, e2 are given		Entity-Destination	0.943
		Cause-Effect	0.928
		Member-Collection	0.911
		Content-Container	0.907
		Message-Topic	0.903
		Entity-Origin	0.897
		Component-Whole	0.878
		Product-Producer	0.873
		Instrument-Agency	0.812
Overall	0.999	Overall	0.899

(a) Original SemEval corpus

NER		RE	
Class	F-score	Relation	F-score
e1,e2,relation mentions are given		range	0.975
		domain	0.963
Overall	0.999	Overall	0.969

(b) New SemEval corpus (OSR) on gold NEs including relation mentions

NER		RE	
Class	F-score	Relation	F-score
Cause-Effect	0.917	domain	0.903
Entity-Destination	0.912		
Member-Collection	0.904		
Message-Topic	0.895		
Entity-Origin	0.883		
Product-Producer	0.874	range	0.902
Content-Container	0.839		
Component-Whole	0.814		
Instrument-Agency	0.807		
Overall	0.877	Overall	**0.902**

(c) New SemEval corpus (OSR) on predicted relation mentions

NER		RE	
Class	F-score	Relation	F-score
N/A		Entity-Destination	0.914
		Member-Collection	0.902
		Cause-Effect	0.888
		Message-Topic	0.883
		Entity-Origin	0.870
		Product-Producer	0.866
		Content-Container	0.834
		Instrument-Agency	0.794
		Component-Whole	0.789
		Overall	**86.33**

(d) After converting predicted OSRs to the original SemEval relations for the comparison purpose

表 2　SemEval 2010 Task 8 に関する評価結果

表 2 に英語の SemEval 2010 Task 8 データに関する評価を示す．こちらは，逆にトータルの性能が低下する結果となった．これは，このデータが交通教則に比べて「Content-Container」のように意味的な関係が薄い関係を扱っているため，関係用語を導入する効果が薄く，また関係用語が in など前置詞になることが多く，関係を表す自立語が存在しない場合が多かったためであると考えられる．

本研究の過程で，オントロジー形式の抽出後，オントロジーに変換する際には，さらに多くの課題が存在することに直面した．今後，関係やイベントをいかにして整合した知識として，オントロジーに加えていくかが大きな研究課題であると考えている．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Bou Savong、Miwa Makoto、Sasaki Yutaka	４．巻 84
２．論文標題 Two evaluations on Ontology-style relation annotations	５．発行年 2024年
３．雑誌名 Computer Speech & Language	６．最初と最後の頁 101569～101569
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.csl.2023.101569	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 相川渉，三輪誠，佐々木裕
２．発表標題 交通に関する知識グラフを用いた運転免許試験問題の解法
３．学会等名 言語処理学会第28回年次大会
４．発表年 2022年

１．発表者名 Savong Bou, Naoki Suzuki, Makoto Miwa and Yutaka Sasaki
２．発表標題 Ontology-Style Relation Annotation: A Case Study
３．学会等名 12th Language Resources and Evaluation Conference (LREC-2020)（国際学会）
４．発表年 2020年

１．発表者名 深谷竜暉，三輪誠，佐々木裕
２．発表標題 タスク指向対話システムの外部表知識の参照能力向上
３．学会等名 言語処理学会第28回年次大会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	三輪　誠 (Miwa Makoto) (00529646)	豊田工業大学・工学部・准教授 (33924)	
研究協力者	Ｂｏｕ　Ｓａｖｏｎｇ (Bou Savong)	豊田工業大学・工学部・研究員	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計1件

国際研究集会	開催年
International Workshop on Symbolic-Neural Learning	2022年～2022年

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関