人工知能学会 第47回SWO研究会
ISWC2018サーベイ

2019年3月10日・11日 沖縄県 石垣島

企画の趣旨

- 本研究会の主な研究領域であるセマンティックウェブのトップカンファレンス**ISWCの最新動向**を，研究会の参加者で分担調査し，その結果を**国内コミュニティで共有**する
- **ISWCへの投稿**と**国内研究コミュニティの活性化**を図る

実施方法

- 本会議のほぼ全セッションについて**1論文1スライド**で紹介する
- スライドは**研究会Webページで公開**する

ISWC2018情報源

- プログラム
  - http://iswc2018.semanticweb.org/program/
- 論文
  - http://iswc2018.semanticweb.org/proceedings/
  - http://ceur-ws.org/Vol-2180/
- 受賞情報
  - http://iswc2018.semanticweb.org/2018/11/09/best-paper-and-best-student-paper-awards-and-nominees/
- プレゼン動画とスライド
  - http://videolectures.net/iswc2018_monterey/

担当者一覧

3月10日(日) 21:00-22:00 ISWC2018サーベイ1

No.	担当者(所属)	担当セッション
1	亀田 尭宙 (京都大学東南アジア地域研究研究所)	Oct10, Session 1.1 Entity Recognition and Linking Oct10, Session 3.2 Information Extraction
2	生島 高裕 (数理先端技術研究所)	Oct10, Session 1.2 Languages Oct10, Session 2.3 Semantics and Truthfulness
3	和田 斉 (NABLAS株式会社、富士山マガジンサービス)	Oct10, Session 1.3 Querying Oct12, Session 1.4 Industry
4	西村 悟史 (産業技術総合研究所)	Oct10, Session 1.4 Industry Oct11, Session 2.3 Healthcare Applications
5	山本 泰智 (ライフサイエンス統合データベースセンター)	Oct10, Session 2.1 Knowledge Access
6	古崎 晃司 (大阪大学)	Oct10, Session 2.2 Ontologies & Data Integration Oct11, Session 1.1 Ontologies & Data Integration

3月11日(月) 14:30-15:30 ISWC2018サーベイ2

No.	担当者(所属)	担当セッション
7	加藤 文彦 (国立情報学研究所)	Oct10, Session 3.3 Scholarly Data Oct11, Session 1.3 Ontologies & Linked Data
8	山口 敦子 (ライフサイエンス統合データベースセンター)	Oct11, Session 1.2 Querying Oct12, Session 1.2 Querying
9	森田 武史 (慶應義塾大学)	Oct11, Session 2.2 Ontologies & Data Integration Oct12, Session 1.1 Ontologies & Data Integration
10	江上 周作 (電気通信大学)	Oct11, Session 2.4 Ontologies & Linked Data
11	Phuc Nguyen (NII)	Oct12, Session 2.1 Embeddings and Deep Learning
12	濱崎 雅弘 (産業技術総合研究所)	Oct12, Session 2.2 Information Extraction Oct12, Session 2.3 Scholarly Data
13	長野 伸一 (株式会社東芝)	Oct10, Session 2.4 Industry Oct10, Session 3.1 Embeddings and Deep Learning

3月10日(日) 21:00-22:00 ISWC2018サーベイ1

No.	担当者(所属)	担当セッション
1	亀田 尭宙 (京都大学東南アジア地域研究研究所)	Oct10, Session 1.1 Entity Recognition and Linking Oct10, Session 3.2 Information Extraction
2	生島 高裕 (数理先端技術研究所)	Oct10, Session 1.2 Languages Oct10, Session 2.3 Semantics and Truthfulness
3	和田 斉 (NABLAS株式会社、富士山マガジンサービス)	Oct10, Session 1.3 Querying Oct12, Session 1.4 Industry
4	西村 悟史 (産業技術総合研究所)	Oct10, Session 1.4 Industry Oct11, Session 2.3 Healthcare Applications
5	山本 泰智 (ライフサイエンス統合データベースセンター)	Oct10, Session 2.1 Knowledge Access
6	古崎 晃司 (大阪大学)	Oct10, Session 2.2 Ontologies & Data Integration Oct11, Session 1.1 Ontologies & Data Integration

Oct10, Session 1.1
Entity Recognition and Linking

Oct10, Session 3.2
Information Extraction

1. 亀田 尭宙

(京都大学東南アジア地域研究研究所)

SIG-SWO 47st : ISWC2018 Survey Session

KAMEDA Akihiro (Kyoto University)

Entity Recognition and Linking

10/10(水)

11:00 – 11:20 RESEARCH

EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs

Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri and Jens Lehmann

11:20 – 11:40 RESEARCH

An Ontology-Driven Probabilistic Soft Logic Approach to Improve NLP Entity Annotations

Marco Rospocher

11:40 – 12:00 RESOURCES

VoxEL: A Benchmark Dataset for Multilingual Entity Linking

Henry Rosales-Méndez, Aidan Hogan and Barbara Poblete

12:00 – 12:20 JOURNAL

Lessons Learnt from the Named Entity rEcognition and Linking (NEEL)

Giuseppe Rizzo, Bianca Pereira, Andrea Varga, Marieke van Erp, Amparo Elizabeth Cano Basave

Information Extraction

10/10(水)

16:00 – 16:20 RESEARCH ★

Enriching Knowledge Bases with Counting QuantifiersSpotlight Paper

Paramita Mirza, Simon Razniewski, Fariz Darari and Gerhard Weikum

16:20 – 16:40 RESEARCH

QA4IE: A Question Answering based Framework for Information Extraction

Lin Qiu, Hao Zhou, Yanru Qu, Weinan Zhang, Suoheng Li, Shu Rong, Dongyu Ru, Lihua Qian, Kewei Tu and Yong Yu

16:40 – 17:00 RESEARCH

Constructing a Recipe Web from Historical Newspapers

Marieke van Erp, Melvin Wevers and Hugo Huurdeman

Entity Recognition and Linking

Session

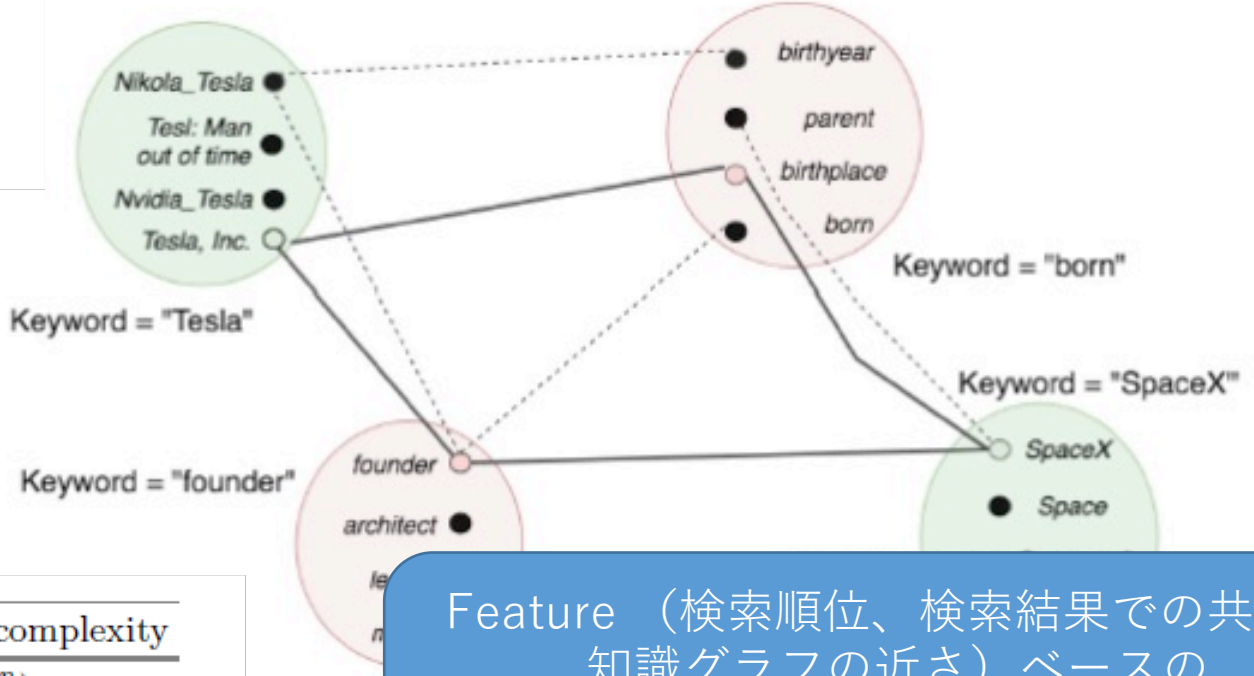
EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs

- Entity と Relation の把握を（直列や別々ではなく）Joint問題として解く *“Where was the founder of Tesla and SpaceX born?”*

$$\mathcal{E} = (1-\alpha) * \mathcal{E}_{BCE} + \alpha * \mathcal{E}_{ED}$$

Binary Cross Entropy
Entity or Relation

RDF2Vecの空間内での
ベクトルの距離



全力のセールスマンと要
領のいいセールスマン

Approach	Accuracy (K = 30)	Accuracy (K = 10)	Time complexity
Brute Force GTSP	0.61	0.62	$\mathcal{O}(n^2 2^n)$
LKH - GTSP	0.59	0.58	$\mathcal{O}(nL^2)$
Connection Density	0.61	0.62	$\mathcal{O}(N^2L^2)$

Feature （検索順位、検索結果での共起、
知識グラフの近さ）ベースの
機械学習がスピードと精度のバランス良い
→２つめ以降の実験は全部これ

An Ontology-Driven Probabilistic Soft Logic Approach to Improve NLP Entity Annotations

- 固有表現抽出とかEntity Linkingとか NLP × Semantic Webの王道だけど、同じ「固有表現」でもタスクによって正解が異なることがある
  - “Lincoln is based in Michigan.”→このLincolnは会社の名前

個人的感想:
普通の曖昧性解消やん…

- タスク横断で使える、Probabilistic Soft Logic (Bach, S.H. et al. 2017)による方法を提案するよ
  - 満たすべき一貫性をルール集合として記述して、それをなるべく満たすようにアノテーション

$$1.2 : \mathsf{WorksFor}(b, c) \,\&\, \mathsf{BossOf}(b, e) \rightarrow \mathsf{WorksFor}(e, c)$$

ルールの重み

Body

≦

Head

$$I(a_1) \wedge I(a_2) = max\{I(a_1) + I(a_2) - 1, 0\}$$
$$I(a_1) \vee I(a_2) = min\{I(a_1) + I(a_2), 1\}$$
$$\neg I(a_1) = 1 - I(a_1)$$

$$f(I) = \frac{1}{Z} \exp\left[-\sum_{r \in R} w_r d(r)^p\right]$$

$d(r) = \max\{0, I(r_b) - I(r_h)\},$

VoxEL: A Benchmark Dataset for Multilingual Entity Linking

- German, English, Spanish, French, Italian
  ヨーロッパ系５か国語のEntity Linkingのデータセットを作った
- ３つくらいは言語横断のデータセットがあるけど
  - Entityの数が言語間で異なる (SemEval, MEANTIME)
  - DBpediaはautoなので汚い

Dataset	$\|D\|$	$\|S\|$	$\|E\|$	Mn	Languages
AIDA/CoNLL-Complete [12]	1393	22,137	34,929	✓	EN
KORE50 [13]	50	50	144	✓	EN
IITB [15]	103	1,781	18,308	✓	EN
ACE2004 [26]	57	-	306	✗	EN
AQUAINT [26]	50	533	727	✗	EN
MSNBC [4]	20	668	747	✗	EN
DBpedia Spotlight [18]	10	58	331	✓	EN
N3-RSS 500 [27]	1	500	1000	✓	EN
Reuters 128 [27]	128	-	881	✓	EN
Wes2015 [36]	331	-	28,586	✓	EN
News-100 [27]	100	-	1656	✓	DE
Thibaudet [1]	1	3,807	2,980	✗	FR
Bergson [1]	1	4,280	380	✗	FR
SemEval 2015 Task 13 [20]	4	137	769	✓	EN,ES,IT
DBpedia Abstracts [2]	39,132	-	505,033	✗	DE,EN,ES,FR,IT,JA,NL
MEANTIME [19]	120	597	2,790	✓	EN,ES,IT,NL
VoxEL_R	15	94	674	✓	DE,EN,ES,FR,IT
VoxEL_S	15	94	204	✓	DE,EN,ES,FR,IT

Lessons Learnt from the Named Entity rEcognition and Linking (NEEL)

Journal Track
35p!!

- 2013年からやってきたTweetを対象とした
  Named Entity Recognition & Entity Linking のコンテストから
  得られた知見を紹介するよ
- たくさんの表がある．Mention Detection, Entity Typing, Candidate SelectionでOff-the-Shelfで使えるものの表とか。

Characteristic	TAC-KBP			ERD	SemEval	W-NUT	NEEL			
	2014	2015	2016	2014	2015	2015	2013	2014	2015	2016
Text	newswire web sites discussion forum posts			web sites search queries	technical manual reports formal discussions	tweets	tweets			
Knowledge Base	Wikipedia	Freebase		Freebase	Babelnet	none	none	DBpedia		
Entity	given by Type			given by KB	given by KB	given by Type	given by Type			
Evaluation	file			API	file	file	file		API	file
	partial end-to-end			end-to-end	end-to-end	end-to-end	end-to-end		partial end-to-end	
Target Conference	TAC			SIGIR	NAACL-HLT	ACL-IJCNLP	WWW			

Information Extraction

Session

Enriching Knowledge Bases with Counting Quantifiers

- “California is divided into 58 counties” みたいな数的知識を抽出したい
- 110種250万件の数的知識の抽出に成功し、これはWikidataにあるそれの128%に相当する。
- 特有の問題
  - 1) トランプの５人の子供が～
    →５人だけとは限らない、これを５人と認識するようにすると多くの数的知識を過少に見積もることになる
  - 2) 「この人には子供が居ない」とはわざわざ言わない、表現に偏りがある。バンドは4人とかが多い、数に偏りがある
  - 3) five, third, twin, a, without… 数を含意する表現というのは意外と多様である

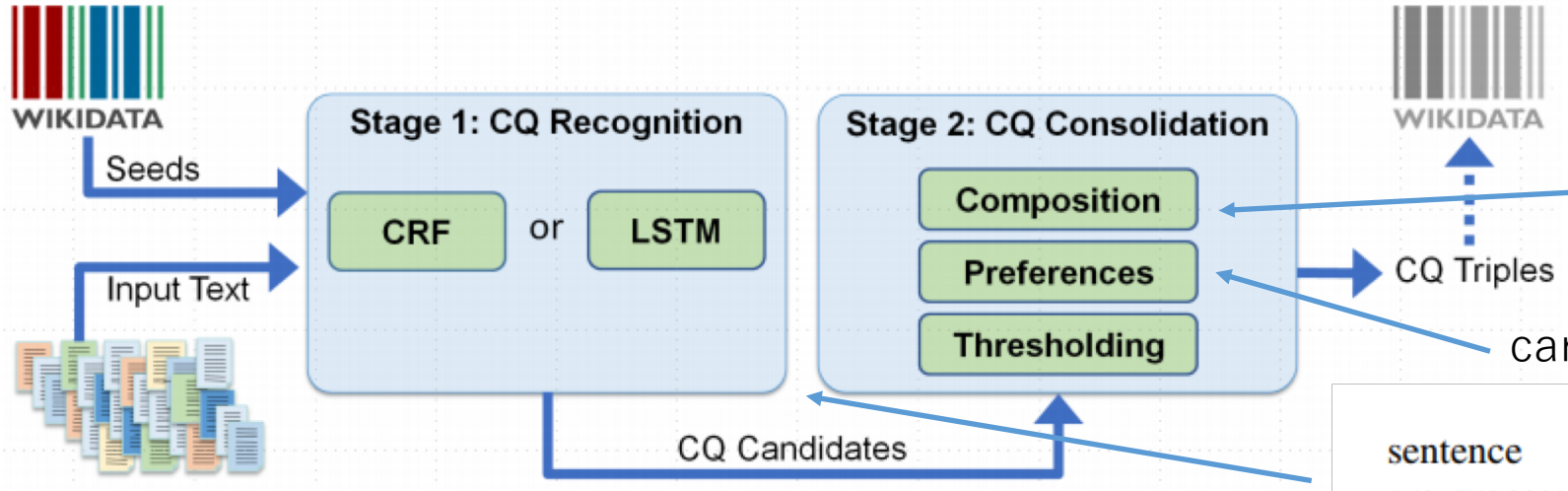


Fig. 2. Overview of the CINEX system.

なぜ…？

Andで結ばれているのを足す

cardinal > numterm > ordinal > article

sentence	*Jolie*	*brought*	*her*	*twins*	,	*one*
pre-processed	*Jolie*	*brought*	*her*	NUMTERM	,	CARDINAL
output tags	O	O	O	COUNT	COMP	COUNT

QA4IE: A Question Answering based Framework for Information Extraction

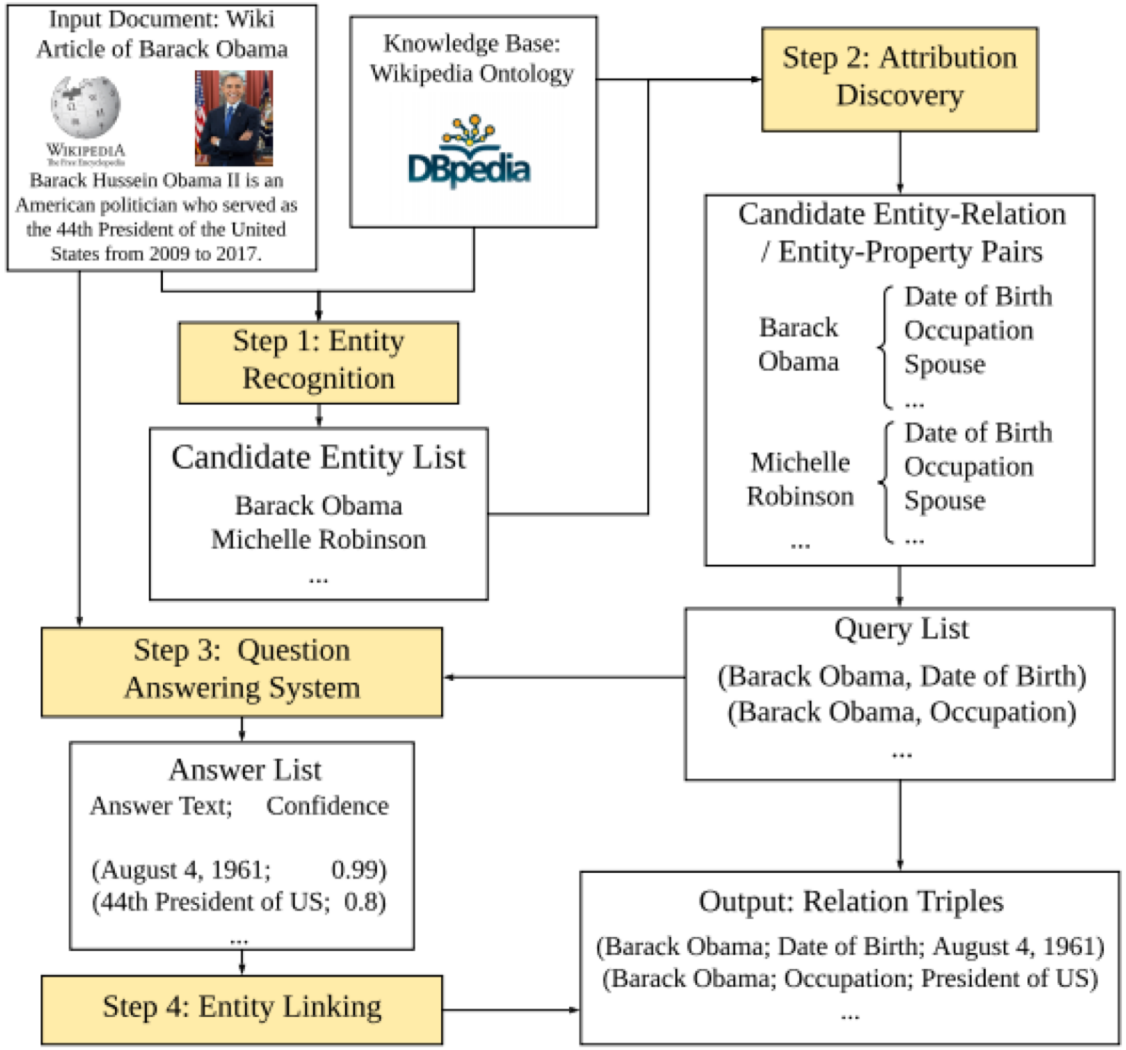


Fig. 1. An overview of our QA4IE Framework.

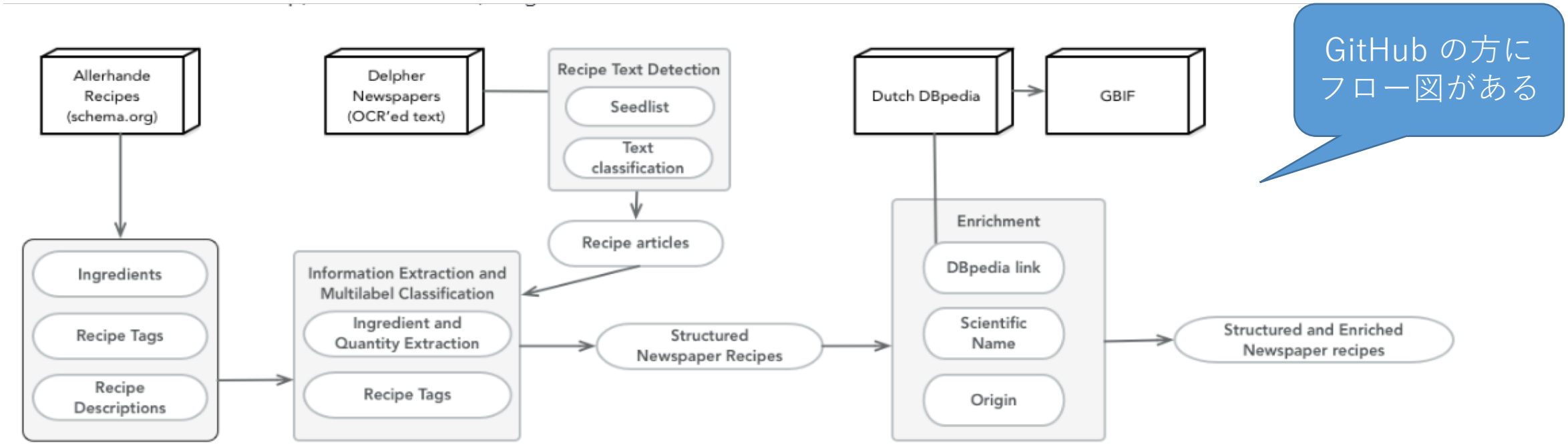
- QA でEntity Linking
  1. 候補集合を作る
  2. 候補集合に関するトリプルをKBから取ってくる
  3. 候補に関するトリプルをもとの文書中からもQA的に抽出する
  4. 2と3の一致を取って、一致したものが正解！

Constructing a Recipe Web from Historical Newspapers

- オランダの古い新聞記事4誌約800万記事をOCRして、掲載されているレシピを分析
- 以下を含むものの30%程度がレシピ記事だった：

‘recept’ or ‘recepten’ *and* one of the following tokens: ‘gram, kilogram, pond, keuken, koken, kook, bakken, eetlepel, gerecht, theelepel, snijden’ (recipe, recipes, gram, kilogram, pound, kitchen, cooking, cook, baking, tablespoon, dish, teaspoon, cut).

- そのアノテーションを使い、unigram, bigramのTF-IDFを特徴にして分類器を学習
  →精度96% Recall95%



Oct10, Session 1.2
Languages

Oct10, Session 2.3
Semantics and Truthfulness

2. 生島 高裕
(数理先端技術研究所)

第47回SWO研究会 ISWC2018サーベイ

1.1.2 Languages

1.2.3 Semantics and Truthfulness

3.2.3 Scholarly Data

株式会社　数理先端技術研究所
2019/3/10 生島高裕

＜目次＞

1) 1.1.2 Languages ＜言語　セッション1.2 研究＞
   a) 19. Semantics and Validation of Recursive SHACL p318-336(19)
   b) 20. Certain Answers for SPARQL with Blank Nodes p337-353(17)
   c) 21. Efficient Handling of SPARQL OPTIONAL for OBDA p354-373(19)
   d) 22. Representativeness of Knowledge Bases with the Generalized Benford’s Law p374-390(17)
2) 1.2.3 Semantics and Truthfulness ＜意味論と真実　セッション2.3 研究＞
   a) 36. Cross-Lingual Classification of Crisis Data p617-633(17)
   b) 37. Measuring Semantic Coherence of a Conversation p634-651(18)
   c) 38. Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage p652-668(17)
   d) 39. Content Based Fake News Detection Using Knowledge Graphs p669-687(19)
3) 3.2.3 Scholarly Data ＜学術データ　セッション2.3=>3.3 リソース＞
   a) The Rijksmuseum Collection as Linked Data p1-10(10)
   b) 12. The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas p187-205(19)
   c) 8. The SPAR Ontologies p119-136(18)

Languages 1：Semantics and Validation of Recursive SHACL

**再帰的SHACL(Shapes Constraint Language：形状制約言語)のセマンティクスと検証

背景：

RDFが普及のため、RDFグラフに制約を指定したり、そのような制約の違反を検出するメカニズムが必要になった

RDFで最も有望なスキーマ言語の1つは、最近のW3C勧告であるSHACL

SHACLの仕様では、再帰的制約に対する検証の問題が残されている

SHACLは設計上、他の制約を参照する制約を優先している

提案：

現在の標準に準拠しながら、**任意の再帰を処理**

SHACLの**コア制約要素の形式的意味論を提案**

セマンティクスを定義することは、本質的に**言語の3つの機能、再帰、任意否定**、および上記で導入されたターゲットベースの**検証メカニズム**の組み合わせが本質

難点の1つは、否定が参照サイクルで任意に使用される、いわゆる**非層別制約による形状の検証を満足のいく方法で定義すること**

グラフ検証は、検証中のグラフ内のノードへのSHACL「形状」の割り当ての存在に基づいており、検証プロセスのターゲットを検証しながら、どの形状が検証または違反されているかを示す

代入が部分的な場合、グラフの一部のノードでの制約の真理値が未知の場合がある場合の検討が必要

SHACL制約に対して**RDFグラフを検証することはグラフのサイズにおいてNP困難**であるため、再帰を扱うことにも代償がある

下限は、層別否定を伴う制約にも当てはまり、**検証問題に対する扱いやすい近似も提案**

構成：

第2章：再帰的SHACL制約検証の問題を**具体的な例**とともに説明

第3章：簡潔な抽象構文とともにSHACLの堅牢な**セマンティクスを定義**し、その正式な**特性**を調べる

第4章：この意味論の下での**グラフ検証問題の計算の複雑さ**を研究

第5章：(検証中のグラフのサイズで）扱いやすさを取り戻すために**健全な近似アルゴリズム**を提案

第6章：RDFに重点を置いて、グラフ制約の検証のための代替言語**形式的な意味論を概説**

継続：

一組の形の（有限の）**充足可能性**、またはSHACL制約の存在下での**SPARQL質問封じ込め**のような他の**問題調査**

この形式化は、任意の再帰を処理するために、ShEx(Shape Expressions)などのグラフの他の制約言語にも拡張できるほど抽象的であることを期待

Languages 2：Certain Answers for SPARQL with Blank Nodes p337-353(17)
空白ノードを含むSPARQLに関する特定の回答

背景：

RDFグラフ内の**空白ノード**は、その識別情報は不明のままで、**ある値を表すために使用**される
例えばBeowulfの作者が空白ノードとして与えられている**Wikidataデータセット**がある
空白ノードの存在するデータに、SPARQLクエリを実行すると、**直観に反する結果**になる可能性がある
不完全な情報は、Web上のデータ管理にとって**大きな課題**
Web上のデータを活用するための**基本的な問題**は、不完全なデータセットに対するクエリの**セマンティクスの定義方法**（および処理方法）
未知の値を持つデータベースに対する、正しい答えを提供するための**クエリ書き換えを提案する最近の研究**がある
RDF標準は、存在的意味論を用いて空白ノードを**明確に定義**
SPARQLは、RDFデータ内の空白ノードの**標準的な存在的意味論**には従わない
意味論のもとでのクエリ評価はかなりの計算オーバーヘッドを招く
「完全なデータベース」の場合、標準のクエリ評価問題のデータの複雑さは**AC0**
未知の値を持つ類似した複雑さはcoNP-hardnessに飛躍

提案：

データ内に未知の値（空白ノード）がある場合、SPARQLクエリの評価問題の再検討提案

貢献：

一般的なSPARQLエンジン（**VirtuosoとFuseki**）の書き換え戦略を評価（基本戦略と書き換えられたクエリのパフォーマンスの比較）
Virtuosoの相対的なパフォーマンス結果は、PostgreSQLに対して公開された結果と競合する

継続：

分析対象の一連のSPARQLクエリを含むデータセットについてはすぐにはわかりませんが、
既存の空白を使用して他のデータセットについて同様の調査を行うことは興味深い

Languages 3：Efficient Handling of SPARQL OPTIONAL for OBDA p354-373(19)
OBDA(Ontology-Based Data Access)用SPARQL OPTIONALの効率的な処理

背景：

OPTIONALは、**欠けている情報を扱う**ためのSPARQLの重要な機能
複雑さでも知られているため、OPTIONALを使用して**効率的にクエリを評価することは困難**
OBDAは、RDFグラフとしてデータの高レベルの表現をユーザーに提供することによって**データベース構造の複雑さ**をユーザーからオントロジーによって**強化されたデータベースに対するビュー**と見なすことができる
LEFT JOINの最適化は、クエリの構造、データソース、およびマッピングに対する「推論」の形式と見なすことができる

提案：

SPARQLクエリに答えるために、OBDAシステムはDBMSによって評価されるがそれを**SQLクエリに再定式化**
（1）SPARQLクエリをRDFグラフの三項関係トリプルにわたって**関係代数式に変換**
（2）**マッピング内のマッチング定義**によってトリプルの発生を**置き換える**ことによって得る（展開）
欠けている情報を処理するために導入された**OPTIONAL演算子を含むSPARQLクエリ**に興味がある
したがって、**RDBのLEFT（OUTER）JOIN演算子と同様の目的**に役立つ
この作業の焦点は、OBDA設定でOPTIONALを使用した照会の効率的な処理
（a）OPTIONALは実際のSPARQLクエリでは非常に頻繁に発生するため、この問題は実際重要
（b）それは**計算の複雑さの原因**：クエリ評価は、OPTIONAL単独によるフラグメントに対しては**Pspace-hard**
（c）エキスパートによるSQLクエリとは異なり、SPARQLクエリのSQL変換は、より複雑な構造を持つより多くのLEFT JOINを持つ傾向があり、DBMSでは最適化に失敗する可能性がある

構成：

第3章：OPTIONALとMINUSを含むSPARQL 1.1のフラグメントを、LEFT JOINとCOALESCEの使用に依存するリレーショナル代数に簡潔に変換
第4章：翻訳と展開から生じる**複雑なLEFT JOINを使ったSQL問い合わせのための最適化手法**を開発
互換性フィルタリダクション（CFR、4.1節）、LEFT JOIN Naturalisation（LJN、4.2節）、CFRとLJNによる、SPARQLの互換性フィルタとCOALESCEは排除（4.3節）、Natural LEFT JOIN Reduction（NLJR、4.4節）、JOIN Transfer （JT、4.5節）、[12]を補完するLEFT JOIN Decomposition（LJD、4.6節）

継続：

我々はこれらの技術をOBDAシステムOntop [4]に実装することに取り組んでいる

Languages 4：Representativeness of Knowledge Bases with the Generalized Benford’s Law
p374-390(17)　一般化されたベンフォードの法則による知識ベースの代表性

背景：

DBpedia、Wikidata、YAGOなどのナレッジベース(KB)には、膨大な数のエンティティと事実が含まれている。
ナレッジベースに関する規則の導入や統計の計算が行われているが**クラウドソーシング**と利用可能なデータベースの**日和見的集積**から構築されているため、**バイアス**されているのが現実である
Semantic Webの明白な成功の1つは巨大な知識ベース(KB)の構築
結論の信頼性は知識ベースの品質、KBが**非常に不完全**
特定のKBがどれほど代表的であるかがわかっていれば、統計を引き出すためにそれを悪用するのが妥当かどうかを知ることができる

提案：

知識ベース内の**関係の代表性を近似**することが目的
関係の事実によって予想される分布を示す**一般化されたベンフォードの法則**を使用し、**KBを実世界の代表とするために追加する必要があるファクトの最小数を計算する**
パラメータ化された法則は、多くの実世界の数値データセットの最初の有効数字で予想される頻度分布を示してる
ベンフォードの法則を、**KBでどれだけのデータが欠けているのかを推定するため**のゴールドスタンダードとして使用する
知識ベースを補完しなければならない事実の最小数を定量化する

貢献：

- 関係が代表的であるための**欠損値の数の下限を計算する方法**を提示
- ある仮定の下では、計算された下限が教師ありと教師なしの両方の文脈において正しいことを証明する
- 方法が、教師なしコンテキストに対しても教師なしコンテキストに対しても有効であることを、実KBに関する実験で示す

特に教師なしの方法では、DBpediaの事実の**63%を監査**でき、代表的な知識ベースであるためには少なくとも**4600万の事実が欠けている**と推定

構成：

第2章：いくつかの関連した研究をレビュー
第3章：代表性の基本概念を紹介
第4章：一般化されたベンフォードの法則に基づいて代表性を近似するための方法を提案
第5章：実験結果

継続：

完全性のために最近行われたように、我々は知識ベースに対する問い合わせの結果を修正するために代表性を考慮したい

Semantics and Truthfulness 1：Cross-Lingual Classification of Crisis Data p617-633(17)
危機データの言語横断的分類

背景：

危機に対する投稿の関連性を分類方法は、異なる言語の投稿を扱うのに苦労しており**急速に進化する危機の状況の間**に**各言語**について**新しいモデルを訓練する**ことは**実行不可能**である

投稿量の多さ:ハリケーンハーヴェイでは、1ヶ月以上で700万以上のツイートが災害について投稿がある

関連の無い:データ専用のハッシュタグを使用しているものでさえも、危機との関連性がほとんどないという問題は悪化している

危機の間に無関係な投稿を迅速に除外し、それによって問題のある投稿のみへの負荷を減らすために、危機情報の関連性を自動的に評価できるツールの必要性がますます高まっている

危機データのさまざまな分類方法が検討されており、教師付き機械学習アプローチを使用している

（Nグラム、テキスト長、POS、ハッシュタグなどの統計機能）

提案：

英語、スペイン語、イタリア語で書かれた記事からなる30の危機的事象からの言語横断データセットについて**統計的および意味的分類アプローチをテスト**する

これまでの研究では、モデルを**ある種類の危機**（例：洪水）について訓練し、それを**別の例**（例：山火事）に適用する際に意味的特徴を追加すると分類精度が向上することを示した

この論文では、モデルが**ある言語**（例えば英語）で訓練されているが、入ってくる記事は**別の言語**（例えばスペイン語）で訓練されているという言語の問題に取り組む

危機情報の関連性を分類するための、教師付きモデルの多言語適合性を高めることに対し、意味を追加することの役割を探る

貢献：

1. **BabelNetとDBpedia**から抽出したセマンティクスを使用して、統計的意味分類モデルを構築
2. **3つの言語**（英語、スペイン語、およびイタリア語）での30の危機イベントからのツイートの関連性を分類することを試みる
3. 言語間のデータセットだけでなく、単一言語に翻訳されたデータセットを使用して関連性分類を実行
4. **セマンティクスを追加**すると、伝統的な統計モデルと比較して、言語間の分類精度が平均1で8.26%-9.07%向上することを示す
5. データセットが同じ言語に翻訳されているときはBabelNetセマンティクスを使用するモデルは、統計モデルよりも3.75%優れている

Semantics and Truthfulness 2：Measuring Semantic Coherence of a Conversation
p634-651(18)　*会話の意味的コヒーレンスの測定

背景：

会話システムの普及は拡大中、インテリジェントな応答のためには、会話の構造とセマンティクスを正しくモデル化する必要がある
要求会話が**首尾一貫**しているということは、可能な対話状態と候補応答に対す**確率分布を改善**することを意味する
会話の本質的な特性はその**「感覚の連続性」**と表現される**一貫性**である

提案：

セマンティックコヒーレンス（意味の位相の揃い具合）

- 認知過程の結果を構成している主観的な尺度
- 参加者の背景知識と常に相対的であり、発話者の解釈に依存
- オブザーバーが概念間の意味のある関係を認識し、導入される新しい関係にクリティカルである能力を反映する
- 意味は、会話の中で提示された知識と、観察者の記憶された世界の知識との相互作用を通して現れる
- 会話には解釈が割り当てられ、且つそれはエージェントに利用可能な知識に依存する
- 対話過程で言及された概念間で成立する関係

セマンティックコヒーレンスは、**発話間の暗黙の関係を推論する**ために膨大な量の異種の知識源を操作すること、すなわち**自然言語を理解する**際の**意味のギャップ**を埋めることを必要とする重要で挑戦的な問題を導入する

貢献：

背景知識の源として**知識グラフ、それらのベクトル空間埋め込みおよび単語埋め込みモデル**を使用して、意味的コヒーレンスを測定するためのグラフベースおよび機械学習ベースのアプローチを提案し、評価する。
Ubuntu Dialogueコーパスでの会話において、これらのアプローチが異なるコヒーレンスパターンをどのようにして発見できるかを示す
（1）対話グラフ表現導入（対話コーパス内の関係を捉えるため）
（2）セマンティックコヒーレンス測定タスクをバイナリ分類タスクとして定式化し、実際の対話と生成された敵対者サンプルを区別
（3）最先端アルゴリズムおよび新規アルゴリズムの性能調査
　ベクトル埋め込みを用いて訓練された**top-k最短経路誘導部分グラフとCNN**
背景知識に関してコヒーレンスを測定することによって会話におけ**意味論的ドリフト**を検出

継続：

グラフ上での**end-to-end学習**など、エンティティリンクの不確実性に対して知識グラフを埋め込んだモデルの堅牢性追及
構造化、非構造化データ両方の証拠を組み合わせることは**知識の希薄性を軽減**し、**サポートと意味関係の解釈可能性**を高めるのでは

Semantics and Truthfulness 3：Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage p652-668(17)真実の発見とRDFの知識ベースを相互の強みに組合わせ

背景：

DBpedia、Yago、Google Knowledge Vaultなどのいくつかの**一般的なイニシアチブ**は**自動的にWebデータを持つ知識ベース（KB）を生成**する
この**ナレッジベースポピュレーション（KBP）プロセス**のパフォーマンスは、KBの品質を保証するために重要
相反するデータが異なるソースから抽出される複雑なケースを扱う
最も頻繁に提供されるデータ値のみを考慮する投票、または単純な戦略に基づくアプローチは事実上制限されている
このようなアプローチでは、Web上で**一般的なスパムベースの攻撃や重複エラーに対処**することができない
信頼性のある情報を自動的に取得するために、**知識融合**に基づくいくつかの技術が提案されている
情報の正確性がソースの**信頼性に厳密に依存**していると考えている

提案：

Truth Discovery (TD:真実発見) パフォーマンス向上

- **目的**：TDは、相反する請求が複数の情報源によって行われた場合に事実（真の請求）を識別すること
- **仮定**：**真の主張は信頼できる情報源**によって提供され、**信頼できる情報源は真の主張**を提供する
- **TDモデル**：どの主張が真実であるかを判断するために、**値の信頼性と情報源の信頼性を繰り返し計算**する
  機能的述語に対して、すなわち現実世界のエンティティの特性に対して**単一の真の値**がある場合に、
  矛盾するトリプルのセットのうち**最も信頼できるものを識別**すること
  ルールマイニングシステムを使用して、これらのパターンを表す**ルールを抽出、教師なし手法**
  証拠は定量化され、推定を改善するために信頼値の計算に統合
  トリプルの正当性を評価するために、使用されるべき**最も有用な規則**を、選択するための方法の提案
  各ルールはその品質に応じてTDパフォーマンスに寄与するため、**既存のルール品質メトリックを集約する関数も定義**している
- **広範な評価**：高品質のRDF KBの作成が信頼性の高いTDモデルの使用からどのように利益を得られるかを指摘

継続：

提案されたモデルの理論的根拠を**他のTDモデルに適用する**ことを計画
他の述語や機能しない述語を考慮することによって、結果を統合するために**評価方法を拡張**すること
ルールがクレームに含まれるものとは異なる値を予測する事実、**否定的な証拠**として考慮していない
情報をどのように取り込むかを、**明示的な公理**、**主観的情報**および**文脈依存性**と一緒に検討することを想定

Semantics and Truthfulness 4：Content Based Fake News Detection Using Knowledge Graphs p669-687(19)　知識グラフを用いたコンテンツベースの偽ニュースの検出

背景：

低コスト、簡単なアクセス、そして急速な普及の特徴を持つ偽のニュースは、**世論を誤解させ、社会秩序を乱し、ソーシャルメディアの信用を傷つけ、当事者の利益を侵害し、そして自信の危機を引き起こす可能**性がある

偽のニュース検出に関するニュース記事を分類するための機能として**執筆スタイル**を捉えることに焦点を当てている

知識ベースのニュースの検出は、事実確認としても知られており、検出はスタイルではなく**コンテンツ**に基づいているため、より有望

既存の**コンテンツベースのアプローチ**は、与えられたトリプルに対して既存の知識グラフでパスを見つけることを試みる**パス到達可能性**に焦点を当ててる

提案：

RQ1：そもそも知識グラフがなく、記事しかない場合はどうなるか？

偽のニュース記事ベースに基づいて**知識グラフを作成**するか、あるいはオープンな知識グラフから関連する**サブグラフを利用**すること

RQ2：偽のニュース検出に不完全で不正確な知識グラフを使用することはできるか？

不完全で不正確な知識グラフに対応するために**意味グラフの類似性を計算**するために**知識グラフの埋め込み**を利用することを考えている

RQ3：コンテンツベースの偽のニュース検出にナレッジグラフ埋め込みをどのように使用できますか？

TransEを利用して、開かれた知識グラフのサブセット、またはいくつかのニュース記事に基づいて構築されたものなど、与えられた知識グラフ上の単一モデルを訓練するアプローチを提案

ネガティブモデルとポジティブシングルモデルを組み合わせた**バイナリTransEモデル**を生成する方法を提案

性能を改善するために、上記の**モデルによって生成された特徴ベクトルを結合**するために融合戦略を使用するハイブリッドアプローチも提案

貢献：

バイナリモデルアプローチが我々の単一モデルアプローチより優れていること

ハイブリッドアプローチが偽のニュース検出の性能を改善すること

テストデータセットの**Knowledge Stream**アプローチよりも優れていること

継続：

（1）コンテンツベースのアプローチとスタイルベースのアプローチ**組み合わせる**

（2）**不完全で不正確な知識グラフ**であっても、結果の説明に偽のニュース検出を提供する

（3）偽のニュースの検出における**知識グラフのスキーマ**の使用、**おおよその推論および不確実な推論**を探求

Scholarly Data 1：The Rijksmuseum Collection as Linked Data p1-10(10)
リンクデータとしての国立美術館コレクション

背景：

多くの美術館は現在、彼らのコレクションへの**オンラインアクセス**を提供している
Linked Dataは、何十年もの間、文化遺産分野におけるデータの**公開と統合を約束**してきた
文化遺産のコレクションをLinked Dataとして公開すると、データの**再利用性が向上し、他のデータソースとの統合が容易**
コレクションアイテムのコンテキストを提供する概念は、多くの場合、複数の**文化遺産組織間で共有**されている
広範囲にわたる関心と明らかな利点にもかかわらず、限られた数の機関しか彼らのコレクションをLinked Dataとして利用できるようにしていない

提案：

Rijksmuseum(アムステルダム国立美術館)コレクションのLinked Dataバージョンには、いくつかの**ユニークな機能**がある
博物館の従業員が**出版プロセスを管理**し、コレクション管理システムからLinked Dataへの**変換レイヤを作成および管理**している
デジタル化プロセスには、**文脈上の概念**を追加するために外部データセット使用し、**手動キュレーションリンクの作成**などがある
データは継続的に拡張され、新しいオブジェクトと説明が追加され、可能な場合はメタデータと画像の両方がオープンライセンスの下で公開

貢献：

博物館はLinked Dataとしての独自のコレクションの出版プロセス全体を管理
オランダ国立美術館は2013年にパブリックドメインで**高画質画像を公開**
オンラインで表示されていたデジタル表現をより詳細に制御し、多くの劣ったバージョンをその高品質の画像**置き換える**ことができた
メタデータの品質と正確さは博物館にとって最も重要
国立美術館には、メタデータの正確性を保証するための広範な**品質管理プロセス**がある
コレクション管理システムに直接変換レイヤを追加することで、同じレベルの品質がLinked Dataバージョンに確実に変換されるようになる
データはRDFで利用可能、構造化された語彙への多くのリンクがあり、コレクションに関する**メタデータ**が利用可能
スミソニアンアメリカ美術館のLinked Data版の概念は、国立美術館のシソーラスにリンクされている
要求される専門知識が利用可能であるならば**機関が彼ら自身のデータを公表することが望ましい**と信じる
最も適切なデータモデル、**URI命名体系、他のデータセットへのリンク、および更新プロセスの選択を管理**続ける
国立美術館のデータは絶えず変更され、新しくデジタル化されたオブジェクトは毎日追加され、従業員は定期的に情報の拡張や詳細化を実行Linked Data版では、作品をより広い文脈の中に置き、**他の人が簡単に再利用できるという進歩**とデータに**新しい視点を加える可能性**がある

Scholarly Data 2：The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas p187-205(19) ***コンピュータサイエンスオントロジー：研究分野の大規模分類

背景：

研究分野のオントロジーは、研究環境を特徴付け、調査し、分析するための重要なツール

研究データの大規模データセットの**統合**、学術的景観の**探索**、科学論文からの情報**抽出**などを容易にするために使用

研究トピックとその関係を記述するオントロジーは、研究のダイナミクスを**理解**するのに役立つ、出版物を**分類**する、研究コミュニティを**特徴付**け、研究動向を**予測**するための非常に貴重なツールです

いくつかの研究分野は、大規模かつ最新の分類法、生物学における**MeSH**および物理学における**PhySH**によってよく説明されている

逆に、現在の**コンピュータサイエンス**の分類法は粒度が粗く、ゆっくり進化する傾向がある

提案：

CSOは、コンピュータサイエンス分野の16M科学論文のデータセットにKlink-2アルゴリズムを適用することによって作成

CSOの2つの主な利点

(i)他の分類には現れない**非常に多数のトピック**が含まれている

(ii)それは最近の出版物のコーパスでKlink-2を走らせることで**自動的に更新**できる

研究テーマのきめ細かい表現は、研究論文の内容を研究者が通常活動する粒度レベルで特徴付けるのに非常に役立つ

CSOは、Linked Data、Semantic Webサービス、Ontology Matching、SPARQL、OWL、SWRL、およびその他40のサブトピックに従ってセマンティックWebを特徴付ける

貢献：

CSOは、Springer Natureの編集チームが現在使用している2つのツール、**Smart Topic Miner**と**Smart Book Recommender**に力を与えている

(1)CSOから作成されたトピックとSNで使用されている内部分類から選択されたタグを使用してSNブックに**注釈**を付けるための半自動ツール

(2)オントロジーに基づく**推薦システム**で、書籍、雑誌、会議録を特定の場所で売り出すことを提案

CSOポータルは、ユーザーがCSOをダウンロードし、探索し、詳細なフィードバックを提供できるようにする**Webアプリケーション**

ポータルは、オントロジーを探索し、トピック間の関係のネットワークを調べるための**3つの異なるインターフェース**を提供

トピックとトピック間の**関係の両方を評価**したり、新しいトピックと関係を提案したりすることもできる

継続：

オントロジーの前バージョンに対するユーザーの**フィードバックの量とセンチメント**を考慮した新しいバージョンのKlink-2を開発している

オントロジーの学習手法を生物学や工学などの他の研究分野にも適用する予定

科学のほとんどの分野を説明する**大規模でデータ駆動型のオントロジーの包括的なセットを作成**すること

Scholarly Data 3：The SPAR Ontologies p119-136(18)

*SPAR（Semantic Publishing and Referencing）オントロジー

背景：

この10年間で、学術出版において新たな進化の一歩が踏み出され**研究情報の出版と共有の方法が劇的に変化**

記事や科学データのような公開された実体を**機械可読にし、発見、閲覧、対話をより簡単にする**ためのWebおよびセマンティックWeb技術の使用この分野の研究者は、**Semantic Publishing**という名前を使用してこの用語を参照しています

提案：

Semantic Publishingの動向

（a）Springer Nature、Elsevier、F1000などの出版社を含む、学界と産業界の両方の人々を積極的に巻き込む

（b）イタリア国立大学評価機関評価機関

（c）広範な研究の質を評価する機関Force11やLinked Researchなどの学術コミュニティはすべて、研究コミュニケーションの現在の慣行を変更し、改善するという目標のもとに結束

より学術的なデータ（例えば、学術出版物のメタデータ、それらに関連する研究データ、および研究で使用される実験および計算のワークフロー）は、学術コミュニケーション分野全体に関わるものが多くなる

恩恵には、視認性の向上、関連研究の**発見可能性、および引用によって明確に関連していなかった作品間の未知の関連性の発見可能性**含まれている

貢献：

学術的成果物のメタデータからワークフローの仕様まで、出版ドメインの主要分野記述に使用できる一連の補完的**直交的なオントロジーの開発**

学術書誌的成果物の出版をもたらす**プロセスの開発**

SPARオントロジー、Semantic Web技術を使って学術出版ドメインを記録することを可能にする一連の**オントロジーモジュールの開発**

継続：

(1)SPARオントロジーの一部としてオントロジー的なアーティファクトを利用可能にすることを検討していることを、学術出版ドメインのスペシャリストオントロジーの追加作成者に提案することで、SPARオントロジーへのコミュニティの関与を広げたい

(2)完全にうまく構成された学術出版物の機械可読な記述のWeb出版を容易にするために、著者や出版社がSPARベースのメタデータの（半）自動生成に使用できるツールを開発、学術出版ドメインのための開発および信頼できるデータモデル

Oct10, Session 1.3
Querying

Oct12, Session 1.4
Industry

3. 和田 斉
(NABLAS株式会社、富士山マガジンサービス)

ISWC2018
Session 1.1.3

Querying

和田 斉

WORQ: **WO**rkload-Driven **R**DF **Q**uery Processing

問題

- クラウド分散型データベースに、joinが多く発生するような RDFクエリを投げると負荷が大きい（network shuffling）
- 従来の工夫はクエリ解決に要したデータを直接キャッシュ

貢献

- joinされやすいRDFトリプルを同パーティションにし network shufflingを減らす
- Bloomフィルタによるjoin縮約（Bloom-join）により 中間join結果から無駄なデータを削除
- 中間join結果をキャッシュ
- Bloomフィルタをindexに使うことでunbound property の検索が効率的になる

Answering Provenance-Aware Queries on RDF Data Cubes under Memory Budgets

問題

- Provenance-Aware Queryを低負荷に実現したい
- 従来技術（Jena）では来歴Entityの制約条件が増え、処理時間が増える

貢献

- Provenance-Aware Caching（PAC）を提案
  - Provenance-Aware RDFグラフの断片化手法
  - OLAPクエリに全体最適化した、断片のキャッシュ方法
- 元のクエリを、named graphとキャッシュ断片を使うようなanalytical queryに変換する方法も併せて提案
- Jena TDBに実装して実証

Bash Datalog: Answering Datalog Queries with Unix Shell Commands

概要

巨大な表形式データ（RDF Tripleも含む）の整形・前処理（select/project/join/union）の工夫

問題

- 専用DBはindexing等に時間がかかる
- ツールや言語の特別な知識が必要

貢献

- DataLogスクリプトを関係代数（Relational Algebra）形式を経由してbashに変換するツールを開発（bashlog）
- その変換ツールをwebサービスとしても提供
- 変換後bashの実行速度においてSOTAに匹敵

Towards Empty Answers in SPARQL: Approximating Querying with RDF Embedding

問題

- SPARQLクエリの結果が0だったとき、クエリのどの部分を変えれば結果が得られるかわからない
- 従来手法は、クエリ条件を緩和し検索範囲を広げる

=> 負荷が増す

貢献

- Entityのrelationを反映しつつRDFをベクトル空間に埋め込む（RDF graph embedding）新手法を開発
- それにより近似解と、近似解を与えるクエリをユーザに提示

ISWC2018
Session 3.1.4

Industry

和田 斉

Use Cases of the Industrial Knowledge Graph at Siemens

- 独自に構築したIndustrial Knowledge Graphの紹介
- ユースケースをプレゼン
  - ビルのDigital Twin
  - 経営リスク管理
  - 製造プロセスの監視
  - 大型工業機械の保守・異常検知
  - 工場／プラントの監視

Populating the FLE Financial Knowledge Graph

問題

- 多くの財務データは非構造化データ => ETL作業負荷
  - PDFの財務報告書
  - 同種のデータであっても、Entity識別子やフォーマットが組織によって異なる
  - データ属性の説明ドキュメントがない

貢献

- 独自技術Dynamic Data Loading（DDL）
  - 機械学習によって自動＆リアルタイムにデータのEntityとDomainを特定し、知識ベースに追加
- DDLを組み込んだプラットフォームを開発
  - 専門家でないユーザでもデータ統合を行える
  - ユーザは対話的にreconsiliation（Entity識別子の調整・提案など）も行える

Knowledge-based Question Answering for DIYers

- DIYのQAシステムを構築した
- EntityとActionについて、非factoid型質問にも答えられる
- どのような推論プロセスによって回答が得られたかも説明可能
- 従来のweb検索よりユーザが効率的に答えに辿り着けることを実証

Amazon Neptune:
Graph Data Management in the Cloud

- Amazon Neptuneの紹介
- Apache TinkerPop/Gremlin と RDF/SPARQL
  どちらもサポート
- AWSエコシステムとして、他のAWSサービス同様の特長
  - 高可用性とスケーラビリティ
  - 自動バックアップ＆復旧
  - Point-in-Timeリカバリ
  - 監視
  - 保存データ暗号化（encryption-at-rest）
  - VPCによるセキュリティ
  - 統合アクセス管理

Oct10, Session 1.4
Industry

Oct11, Session 2.3
Healthcare Applications

4.西村 悟史
(産業技術総合研究所)

1.1.4 Industry
2.2.3 Healthcare Applications

西村悟史

1.1.4 Industry

- Knowledge Graph Solutions in Healthcare for Improved Clinical Outcomes
  - Jans Aasman and Parsa Mirhaji
  - Affiliation: Franz Inc, Motefiore Mecidal Center
- Elsevier's Healthcare Knowledge Graph and the Case for Enterprise Level Linked Data Standards
  - Alex DeJong, Radmila Bord, Will Dowling, Rinke Hoekstra, Ryan Moquin, Charlie O, Mevan Samarasinghe, Paul Snyder, Craig Stanley, Anna Tordai, Michael Trefry, and Paul Groth
  - Affiliation: Elsevier
- Accelerating Drug Discovery in Rare and Complex Diseases
  - Shima Dastgheib, Craig Webb, Qiaonan Duan, Rowan Copley, Gini Deshpande and Asim Siddiqui
  - Affiliation: NuMedii, Inc.
- Using Knowledge Graph to improve enterprise search experience
  - Dmytro Dolgopolov and Elena Romanova
  - Affiliation: FINRA (Financial Industry Regulatory Authority)
- Supporting Predictive Models Results Interpretation for Comfortable Workplaces
  - Iker Esnaola-Gonzalez, Jesús Bermúdez, Izaskun Fernandez and Aitor Arnaiz
  - Affiliation: IK4-TEKNIKER

Knowledge Graph Solutions in Healthcare for Improved Clinical Outcomes

- 背景
  - 医療問題：コストを削減せよという政府からのプレッシャー，多様な患者集団に対する超厳しいガイドライン，disruptive technology？
  - 患者理解のためには複雑な因子の掛け合わせの情報が必要だし，いくつかは診療現場で手に入れることは難しい．
    - 例：家とか職場環境，栄養，遺伝子
  - この産業は患者データを長い間集めてきた．
    - データは，自由記述文からセンシング装置からの画像とか波形などさまざまである．
    - 適切なデータにたどり着くことは大事なんだけど，このデータはしばしばすぐには分析できていなかったり，製薬会社とかからの外部データとの結合がされていなかったりする．
- 解決法：PALM: Patient centered Analytic Learning Machine
  - リスクのある患者に旗を立てるとか，医師が最適な治療計画を立てるのを助ける．
  - 色々なデータをプラットフォームが統合している．
    - 知識グラフは情報検索の能力を提供する．
    - 効率的にデータのパターンを見つけるとか，それらのパターンを診療アウトカムを向上させる目的に使う．
  - ある事例では，長期間人工呼吸器を付けている患者から48時間以内に致命的な呼吸器異常が起こる可能性が70%以上の事例を正確に予測した．

Elsevier's Healthcare Knowledge Graph and the Case for Enterprise Level Linked Data Standards

- 背景
  - 組織内の様々な部署から集められたデータは資産と見なされる．
  - Elsevierでは，Health systemの異なる部分に焦点を当てて製品を提供している．
    - 例えば，Order Setsは診断後の患者に対する事前に決められた行為群を与えるし，
    - Clinical Keyは外科医に鍵となる参考情報を与え，
    - Gold Standard Drug Databaseは薬についての包括的で最新の情報を与える．
- 課題
  - 新しい意思決定支援の製品を提供しようと思ったら，これらの既存製品間をまたいだデータアクセスが必須となる．
  - しかし，そのようなデータネットワークを作るのは難しい．
    - １．データは異なるデータベースにあって，APIs (Application Programming Interfaces)は製品ごとに異なっている．
    - ２．これらの製品はデータ集めのpipelineが異なっている．
    - ３．異なるsoftware stacks（.NETからJavaまで）で製品は造られている．共通プラットフォームの上で書き換えるのはコストが高い．
- 解決法
  - 異なるシステムから得られるhealth dataに，開発者がアクセス可能にすることを目的として，Semantic web technologyを使う．
- 教訓
  - １．API centric approach は重要
  - ２．開発サイクルに取り込むために既存の開発者ツールを使うことは重要
  - ３．Semantic web technologyのどこを使うかを決めるのは重要
  - ４．内部の例を提供することは適用を容易にする
  - ５．プラットフォームにとらわれないという性質が，企業内の異質な環境に役に立つ．

Accelerating Drug Discovery in Rare and Complex Diseases

- 背景
  - 慢性疾患の予見や診断，治療や管理に関して著しい発展があるにも関わらず，いまだに，治ることもなく管理を続けなければならない複雑怪奇な状態は多く残っている．
  - 医学生物学のデータはたくさん作られて飛躍的に蓄積している．
    - 今日，個別の細胞からDNAシークエンスの情報を取り出すことも可能になっており，一つの細胞からとてもたくさんのデータポイントが生成されている．
  - 医学生物学のリポジトリにある非常に貴重なデータは新しい治療の選択肢を提供する可能性を秘めている．
- 課題
  - 鍵となる問題は，利用可能なデータの量と多様性とそれらの複雑な関係を考慮した上で意味を解釈することである．
- 解決法
  - NuMediiは異なる開発ステージにある疾患と薬高品質のデータへのアクセスを提供する．
    - Semantic Knowledge Baseを作って，様々な公共や私企業の持つデータソースから異質なデータを集めお互いの意味ある関係を構築し，統合した．
    - この知識ベースはSPARQLで検索できる．
    - OntoText GraphDBを使ってる．GUIも作った．それはドメインエキスパートが探すのを助ける．
- Case Study
  - 突発性肺線維症は治ることはなく，pirfenidoneとnintedanibで繊維化を遅らせることが最良の治療法である．
  - 他の繊維症のDBを組み込んで，700,000以上のPubMed Abstractを抽出した．

Using Knowledge Graph to improve enterprise search experience

- 背景：broker-dealer産業の規制にたくさんの文書を扱う
  - FINRAは規制をする組織
  - 内部システムに大量の非構造化，構造化データを持っている．
  - 投資家も検査する人も分析する人も多さに閉口している．
- 課題：文書のコピーと類似品が多すぎる
  - 検査とか調査の準備にスタッフは何日もかけている．
  - FINRAは内部と外部の色々なシステムから集められたデータセットの中をナビゲートする解決策が欲しい．
- 解決策：Semantic Web, Text Mining, Enterprise Search and Graph Analysisをくっつけて，FINRA Knowledge Graphを作る．
  - Scalable semantic inference engineを作った．人間みたいな推論をして新しいfactを得る．
  - FINRA’s ontologyはscema.org ontologyを拡張して作っている．
  - Text miningはフリーコメントから特徴とかindividualを抽出してデータをリッチにする．
  - Machine Learning modelを使って関連データ要素をクラスタリングして，同じ情報の文書を見つける．
  - Enterprise SearchはSPARQLで強化されて質が上がる．
  - Community detection algorithmが悪人を見つける．
- 規制作業の効率が上がったよ．

Supporting Predictive Models Results Interpretation for Comfortable Workplaces

- 背景
  - 90%の人はほとんど建物内で過ごす.
  - 温度環境が悪いと病気になるし仕事のパフォーマンスが下がる.
- 課題
  - HVAC (Heating, Ventilation and Air Conditioning)が必要.
- やったこと
  - EEPSA (Energy Efficiency Prediction Semantic Assistant)がKDD (Knowledge Discovery in Database) processを通してHVACのControlを支援する.
  - 建物，センサー，デバイスの起動，それらに関連して見られる情報を記述するためのEEPSA ontologyを作りました.
  - EROSO (thermal cOmfort SOlution)はKDD processとSemantic Technologyを合わせるフレームワークです．とくに，KDDの解釈に説明を与えるのにSemantic Technologyが役立つ.
    - 制御戦略に応じて温度予測をする．温度予測に対してオントロジーを用いたannotationがされる．予め決められたルールで予測結果が分類され，施設管理者に提示される．施設管理者はGUIを通じて制約を設定する．設定内容はSPARQLとして温度予測の結果を取り出しBuilding Management Systemに制御データが送られる.
  - 200人規模のオフィスに実装した．@IK4TEKNIKER building in Spain.

2.2.3 Healthcare Applications

- A Systematic Analysis of Term Reuse and Term Overlap across Biomedical Ontologies
  - Maulik R. Kamdar, Tania Tudorache, Mark A. Musen
  - Affiliation: Stanford Center for Biomedical Informatics Research
- Semantic Technologies for Healthy Lifestyle Monitoring
  - Mauro Dragoni, Marco Rospocher, Tania Bailoni, Rosa Maimone and Claudio Eccher
  - Affiliation: Fondazione Bruno Kessler
- Supporting Digital Healthcare Services Using Semantic Web Technologies
  - Gintaras Barisevicius, Martin Coste, David Geleta, Damir Juric, Mohammad Khodadadi, Giorgos Stoilos and Ilya Zaihrayeu
  - Affiliation: Babylon Health
- Analyzing user interactions with biomedical ontologies: A visual perspective
  - Maulik R. Kamdar, Simon Walk, Tania Tudorache, Mark A. Musen
  - Affiliation: Stanford Center for Biomedical Informatics Research

Semantic Technologies for Healthy Lifestyle Monitoring

- 背景
  - 健康な生活様式は慢性疾患の予防に効果的である.
  - モバイルとかウェアラブルデバイスとかでデータを取る技術が発展している.
- 課題
  - データを統合して，個人適応的なアドバイスをして生活様式を変えさせるのは出来ていない.
- 解決法：PerKApp platformによるpersonalized healthy lifestyle recommendations and advicesの提供.
  - (i) 健康な生活様式に関する助言と推奨を生み出す責任を負う知識の構成
  - (ii) システムから得られる知識をモデル化するためのドメインエキスパートと知識工学者の協調環境
  - (iii) 実環境でプラットフォーム全体がどのように振る舞ったかの観測結果
- 結果
  - we evaluated the reasoner performances by discussing:
  - (i) 異なる設定でのユーザデータの処理に係る時間
  - (ii) システムのスループット？（PerKApp platformを異なる環境で動かすときの可能なハードウェア要求を推論するため）
    - the throughput of the system with the aim of inferring possible hardware requirements for deploying the PerKApp platform into different environments; and
  - (iii) ユーザのプロファイル（モニタリングルールの数とか）の複雑さがどのようにシステムの効率全体に影響を与えるか.
  - Key to Health projectとやらの最初の45日間のデータを使って想定シナリオでシミュレーションを行った.

Supporting Digital Healthcare Services Using Semantic Web Technologies

- 背景
  - Babylon Health はユーザにデータ登録をしてもらって医者によるビデオコンサル（薬の処方，専門医の照会，検査の予約）を提供している．
  - 症候チェックとトリアージをchatbotが提供する．
  - テキスト処理で患者の入力したテキストに適したアノテーションを行う．
- 課題
  - BioPortalやBio2RDFがあるが，診療データは失われている．
  - OWLのmodelは複雑すぎるので，エンドユーザには扱えない．
  - 既存のreasonerは厳格すぎる．
- 方法
  - 単純化したモデルを用いる．
    - Clinical Knowledgeと呼ぶ抽象レイヤーを作った．
  - 現実の問題解決のために，あいまいさを扱う．
- Lessons Learned and Conclusions
  - Main lesson learned is “Linked Data Is Merely More Data.”
    - An upper-level-ontology, integrity constraints to check validity of data, building abstraction layers on top of SPARQLが重要．
    - それでも大きく不均一なデータのため，構造の一貫性を保ち包括することは難しい．
  - 良かった点：
    - ほとんど全ての医学的状況を動的に記述することができた．
    - クラス表現の比較に形式意味論を用いたことで高いprecisionを実現した．
    - NLP(Natural Language Processing)を用いることでML-based？よりもrecallを高めることが出来た．
  - Future plan：最初からLDGを作り直す
    - 少ないソースから作って，徐々に大きくするアプローチをとる．

Oct10, Session 2.1

Knowledge Access

5. 山本 泰智
(ライフサイエンス統合データベースセンター)

G. Troullinou et al., "Exploring RDF/S KBs Using Summaries"

課題：複雑なオントロジーを利用者が自由に把握する手法が希少

提案：重要部分抽出とそれに基づく意味的ズームイン・アウトなどを可能に

新規：利用者が自由にズームイン・アウト、そして特定概念を中心とした深堀を可能にする手法とツールRDFDigest+の開発

手法概要：

- ノードの重要度を取得するにはBetweenness、PageRank、HITSの各アルゴリズムを利用。
- 特定のノードについて深堀するためにシュタイナー木アルゴリズムの近似法としてCHINSとShortest Pathsアルゴリズムを適用して結果を比較。
- ズームアウトとインはCHINS/Shortest Pathsにより得られた結果を利用して、重要度に応じて抽出するノード数を増減させることで実現。
- DBpediaへのSPARQLクエリによる評価でHITS/Betweennessが結果良好。
- RDFDigest+公開サイトは次の通り。http://rdfdigest.ics.forth.gr/

M. Atzeni et al., "What is the cube root of 27? Question Answering over CodeOntologySpotlight Paper"

課題：自然言語での質問に対する回答がプログラムコードとその実行は困難

提案：CodeOntologyを構築して、自然言語をSPARQLに変換、結果を得る

新規：オブジェクト指向コードのモデル化とJavaコードのパーサー開発

手法概要：

- 簡素な回答と丁寧な回答を返す手法を提案。
  - 簡素な回答は質問とメソッド記述などとの類似度を見て回答生成。
    → 一つの質問に対して一つのメソッドのみを起動。
  - 丁寧な回答は係り受け構造解析を含む質問解析を行い回答生成。
    → 複数のメソッドの組み合わせを可能として複雑な質問に対応。
- 文字列類似度判定は編集距離やnグラム重複、DBpediaリンクやコサイン類似度を利用。
- 評価方法はStackOverflowでの質疑とWolframAlphaとの比較して提案手法の優位点を主張
- 関連データは次のサイトから。

* http://codeontology.org/
* https://doi.org/10.6084/m9.figshare.607166

J. Moreno-Vega et al., "GraFa: Scalable Faceted Browsing for RDF Graphs"

課題：エンドユーザーが使いやすいRDF検索ツールがない

提案：大規模データにも対応できるファセット検索ツールの開発

新規：ファセット検索高速化のための事前索引データ構築手法

手法概要：

- キーワードもしくは型が与えられることで検索を開始するツールとして予め特定のRDFデータについてLuceneの索引を構築しておく。
  - 対象はWikidata全体で、言語タグ付きリテラルは英語とスペイン語。
  - 1億95百万トリプル、うち1億2千万プロパティ(p,o)、75百万ラベル。
  - キーワード検索結果として5万件以上の主語がヒットするデータを予め索引付け。それを取得するためのSPARQLクエリ数は141。
- 結果として索引付けされたデータがヒットする場合は全て1秒未満で検索。
- 利用者テストを実施したところ、Wikidataの提供するクエリIFの方に軍配。
- 改善してWikidata MLで評価を募った結果、それなりに良い評価をもらう。
- 関連情報公開サイトはこちら。 https://github.com/joseignm/GraFa/

G. D. Tommaso et al., "Wiki-MID: a very large Multi-domain Interests Dataset of Twitter users with mappings to Wikipedia"

課題：自由に使える推薦システム評価用データセットが希少

提案：ツイッターのデータから多分野のデータセットを自動生成

新規：多言語 (英語、イタリア語)、多分野 (90/利用者)、Wikipediaとの連結

手法概要：

- 幅広い分野に対応するために、Wikipediaの情報を活用。
- ツイッターのプロフィール欄やフォローアカウント、特定の記述を用いて、各ツイッター利用者の興味の対象を抽出し、それに関連するWikipediaの記事を特定し、分野を割り当て。
- SIOCとSKOS語彙を用いてRDFデータセットを構築、公開。
- データセットや関連情報の公開サイトは以下の通り。
  http://wikimid.tweets.di.uniroma1.it/wikimid/

Oct10, Session 2.2
Ontologies & Data Integration

Oct11, Session 1.1
Ontologies & Data Integration

6.古崎 晃司
(大阪大学)

2019/3/10 第47回SWO研究会

ISWC2018サーベイ

1.2.2 Ontologies & Data Integration
2.1.1 Ontologies & Data Integration

大阪大学産業科学研究所
古崎　晃司

Oct.10
Session 1.2.2

Ontologies & Data Integration

IN-USE	Synthesizing Knowledge Graphs from web sources with the MINTE+ framework *Diego Collarana, Mikhail Galkin, Maria-Esther Vidal, Christoph Lange, Sören Auer and Simon Scerri*
JOURNAL	Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning *Freddy Lecue, Jiewen Wu*
RESOURCES	A Complex Alignment Benchmark: GeoLink Dataset *Lu Zhou, Michelle Cheatham, Adila A. Krisnadhi and Pascal Hitzler*
RESOURCES	Audio Commons ontology: a data model for an audio content ecosystem *Miguel Ceriani and Gyorgy Fazekas*

1.2.2 Ontologies & Data Integration

IN-USE

Synthesizing Knowledge Graphs from web sources with the MINTE+ framework

Diego Collarana, Mikhail Galkin, Maria-Esther Vidal, Christoph Lange, Sören Auer and Simon Scerri

- **Webリソースを用いてKGを統合するMINTE+フレームワークを， Law Enforcement, Job Market Analysis, Manufacturingの３ドメインに適用.**
- **MINTE+フレームワークの概要**
  1. 様々なWeb APIでRDFのMoleculesを取得
  2. RDF Moleculesの同一性を判定
  3. Semantic Similarityにそって 統合

  ※各ステップで用いるAPIや閾値をパラメータとして設定
- **パラメータの設定例**

Table 1. MINTE$^+$ Configuration. The Law Enforcement Application

Parameter	Value	Description
Query	Free Text	Usually people, organizations, or products name or description
Ontology	LiDaKrA	The ontology describing the main concepts in the crime investigation domain
Web APIs	11	Facebook, Google+, VK, Twitter, Xing, ICIJ Offshore Leaks, DBpedia, eBay, darknet sites, crawled darknet markets, OCCRP reports
Simf	GADES [8]	A semantic similarity measure for entities in knowledge graphs
Threshold	0.9	Only highly similar molecules are synthesized
Fusion policy	Union	No information is lost, e.g., all alias names of a person are kept in the final molecule

Table 3. MINTE$^+$ Configuration. The manufacturing application

Parameter	Value	Description
Query	Provider metadata	Includes company name, address, web site
Ontology	Schema.org	An extension of organization concept is used to describe the providers
Web APIs	4	DBpedia, Google Knowledge Graph, plus further confidential sources
Simf	Silk	Wikipedia page is used in the linking rule
Threshold	1.0	Providers with same Wikipedia page are integrated
Fusion Policy	Authoritative	DBpedia is defined as the main source

1.2.2 Ontologies & Data Integration

JOURNAL **Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning**

Freddy Lecue, Jiewen Wu

- **Semantic webと機械学習の組み合わせにより，大規模な組織における従業員による異常な費用請求**(abnormal expenses claim)**の (i) 確認**(identifying), **(ii)説明**(explaining)，**(iii) 予測**(predicting)**を行う．**
- 旅費を対象．Accenture従業**191,346人を対象**に試用実験．

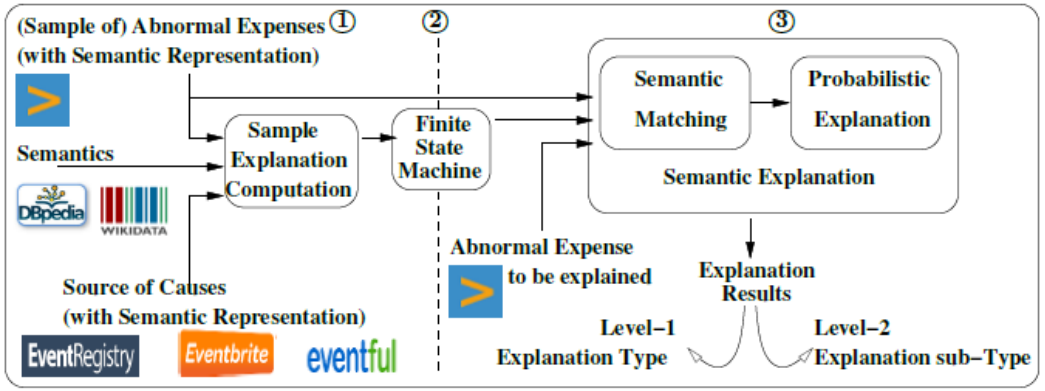
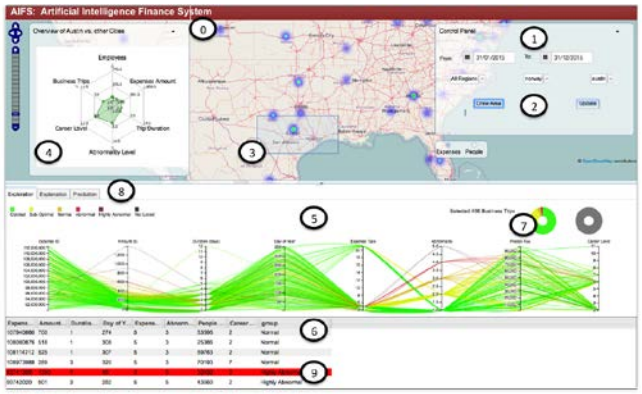
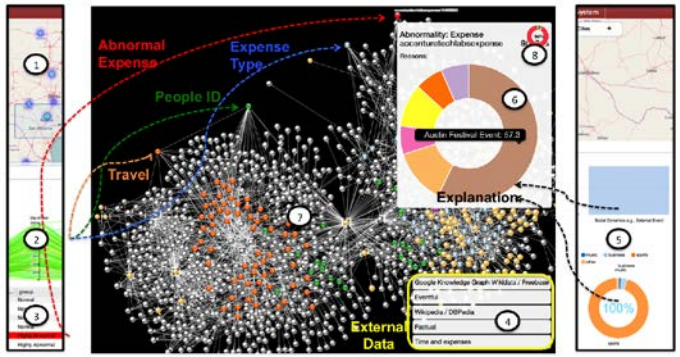


Figure 6: Semantic Explanation Approach Overview. ①: Compilation of (sample of) abnormal expenses and their explanation. ②: Finite State Machine based representation of sample expenses and their context. ③: Semantics-based comparison, evaluation of explanation.





RESOURCES

A Complex Alignment Benchmark :GeoLink Dataset

Lu Zhou, Michelle Cheatham, Adila A. Krisnadhi and Pascal Hitzler

- **複雑なオントロジーアライメントのベンチマーク**
- **対象**：GeoLink Base Ontology (GBO)とGeoLink Modular Ontology (GMO)

Table.1 各オントロジーの規模

Ontology	Classes	Object Properties	Data Properties
GeoLink Base Ontology	40	149	49
GeoLink Modular Ontology	156	124	46

- **複数機関の専門家による手動アライメント**
- **アライメントの種類**
  - 同等（equivalence）1:1, 1:n, m:n
  - 包摂（subsumption correspondences）
- **アライメントの表現**
  - Expressive and Declarative Ontology Alignment Language (EDOAL)
  - Rule syntax.

Table.2 アライメントのパターン

Pattern	Occurrences	Category
Class Equivalence	10	1:1
Class Subsumption	2	1:1
Property Equivalence	7	1:1
Property Equivalence Inverse	5	1:1
Class Typecasting Equivalence	4	1:n
Class Typecasting Subsumption	1	1:n
Property Typecasting Subsumption	5	1:n
Property Typecasting Subsumption Inverse	5	1:n
Typed Property Chain Equivalence	26	m:n
Typed Property Chain Equivalence Inverse	17	m:n
Typed Property Chain Subsumption	17	m:n
Typed Property Chain Subsumption Inverse	12	m:n

RESOURCES

Audio Commons ontology: a data model for an audio content ecosystem

Miguel Ceriani and Gyorgy Fazekas

- **Audioのオントロジー**
  - Repositories や clientsの統合を促進する.
  - 包括的なモデル．既存オントロジーともマッピング.
- **構築方法**
  - **知識獲得**
    - 事前アンケート
      質問数24，回答数661
  - **仕様策定**
    - Use Cases
    - Competency Questions
    - Formal Requirements

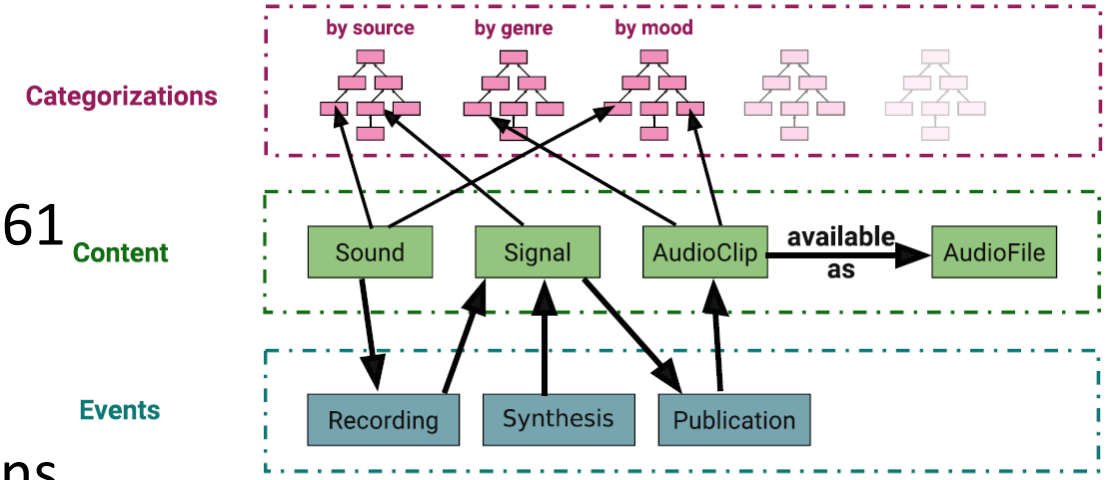


Fig. 1. A layered view of the Audio Commons ontology

- The Audio Commons Ontology https://w3id.org/ac-ontology/aco

Oct.11
Session 2.1.1

Ontologies & Data Integration

RESOURCES Drammar: a comprehensive ontological resource on drama
Vincenzo Lombardo, Rossana Damiano and Antonio Pizzo

RESOURCES DOREMUS: A Graph of Linked Musical Works
Manel Achichi, Pasquale Lisena, Konstantin Todorov, Raphaël Troncy and Jean Delahousse

RESOURCES HeLiS: An Ontology For Supporting Healthy Lifestyles
Mauro Dragoni, Tania Bailoni, Rosa Maimone and Claudio Eccher

RESOURCES Knowledge Integration for Disease Characterization: A Breast Cancer Example
Oshani Seneviratne, Sabbir Rashid, Shruthi Chari, Jim McCusker, Kristin Bennett, James Hendler and Deborah McGuinness

RESOURCES

Drammar: a comprehensive ontological resource on drama

Vincenzo Lombardo, Rossana Damiano and Antonio Pizzo

- Action, Scene, Timeline,..といったドラマの内容を表現するためのオントロジー
- 構築方法論にはNeOnを採用
- Emotionsの計算，シーンの可視化，インテンションの分析などのアプリを想定.
- Webアノテーションツールを用意しており，コーパスやツールは近日公開予定
- http://www.di.unito.it/~vincenzo/ontologies/DrammarOntology.html

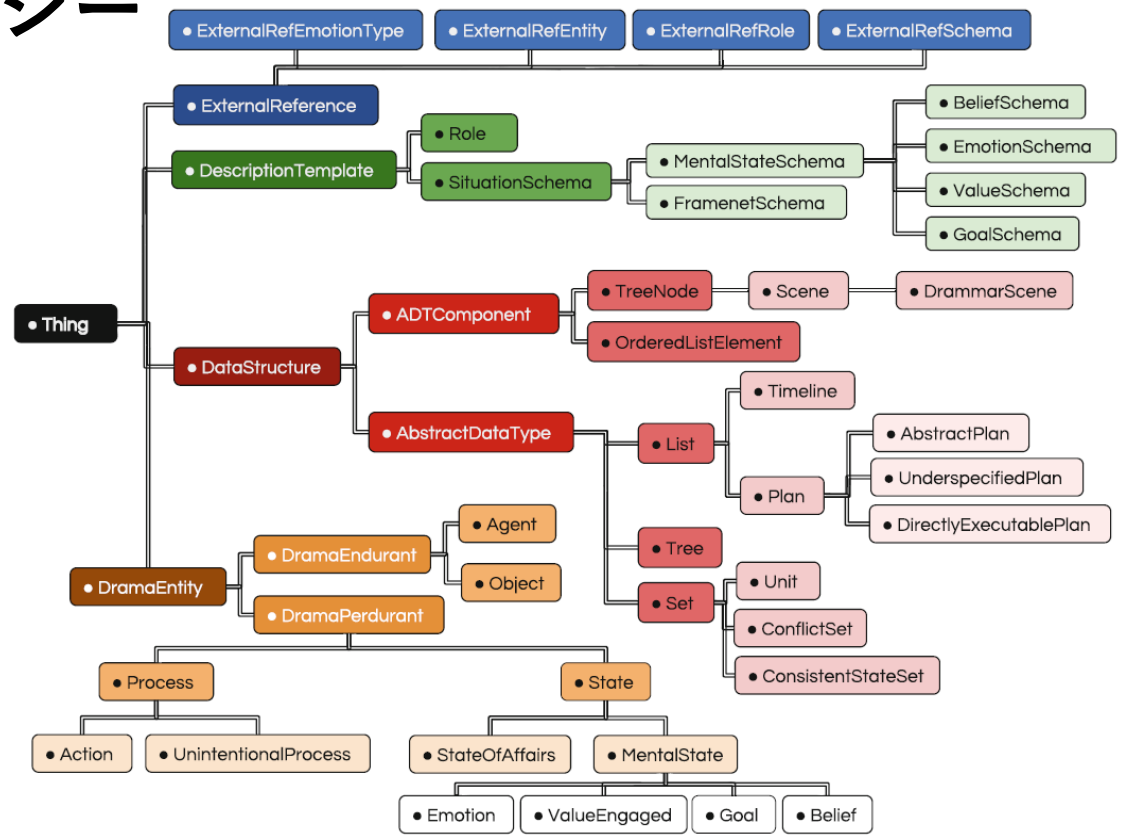


Fig. 1. Taxonomy of ontology Drammar. Colors distinguish sections of the taxonomy (four major sections); box colors desaturate while going to specific classes. (Color figure online)

RESOURCES

DOREMUS: A Graph of Linked Musical Works

Manel Achichi, Pasquale Lisena, Konstantin Todorov, Raphaël Troncy and Jean Delahousse

音楽作品のナレッジグラフ

元データ

- (1) Works and Artists(MARC形式): フランス国立図書館(BNF)
- (2) Works and Concerts (MARC形式) : パリフィルハーモニー(PP)
- (3) Concerts and Recordings(XML形式):ラジオフランス(RF)

再利用したモデル

- CIDOC-CRM
- FRBRoo

ツール

- データのコンバータ
- マッピング
- 検索エンジン
- データ検証　…etc.

リンク数：7945

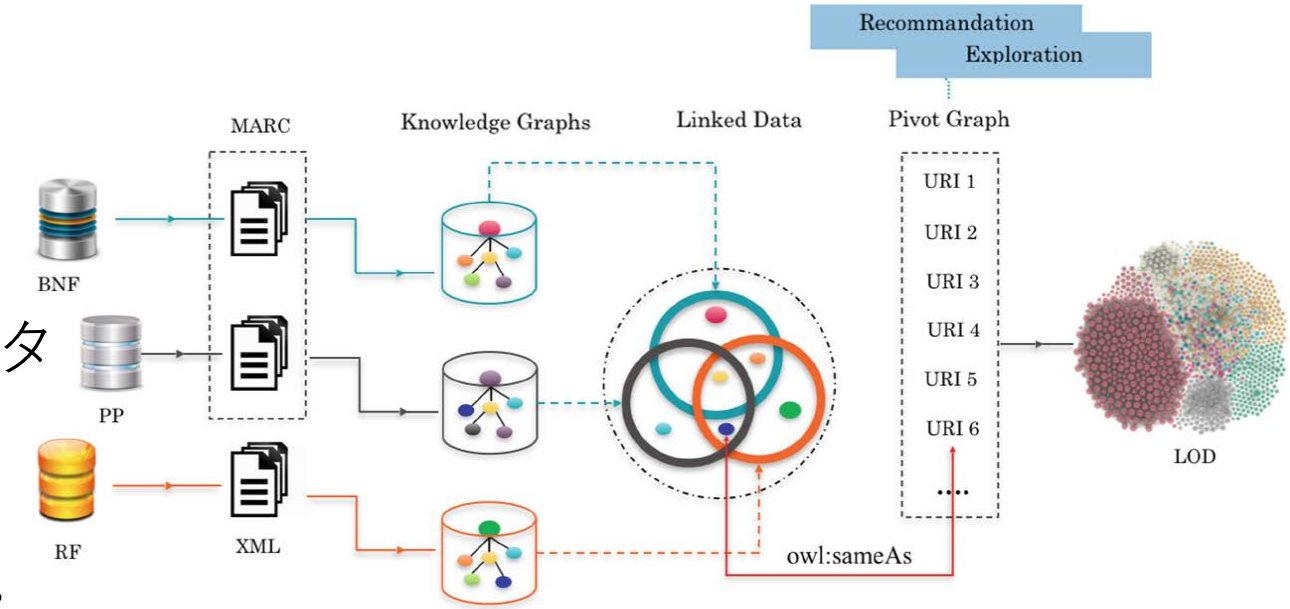


Fig. 1. The DOREMUS data lifecycle.

RESOURCES

HeLiS: An Ontology For Supporting Healthy Lifestyles

Mauro Dragoni, Tania Bailoni, Rosa Maimone and Claudio Eccher

- **食べ物と身体活動のオントロジー．ユーザの不健康な振る舞いをモニタリングするための表現を提供.**
- METHONTOLOGY方法論を利用．非構造化リソースから専門家と共同構築.
- 食品の栄養，イタリア農業省のアーカイブ，レシピDBなど利用.
- AGROVOCとリンクあり.

主要概念

- Food and Nutrient(栄養素)
- Activity：
- Temporal Event
- User Event
- Monitoring Entity

例）ユーザが〇〇を食べたら
栄養素をどの程度得て…
といったモデルを記述

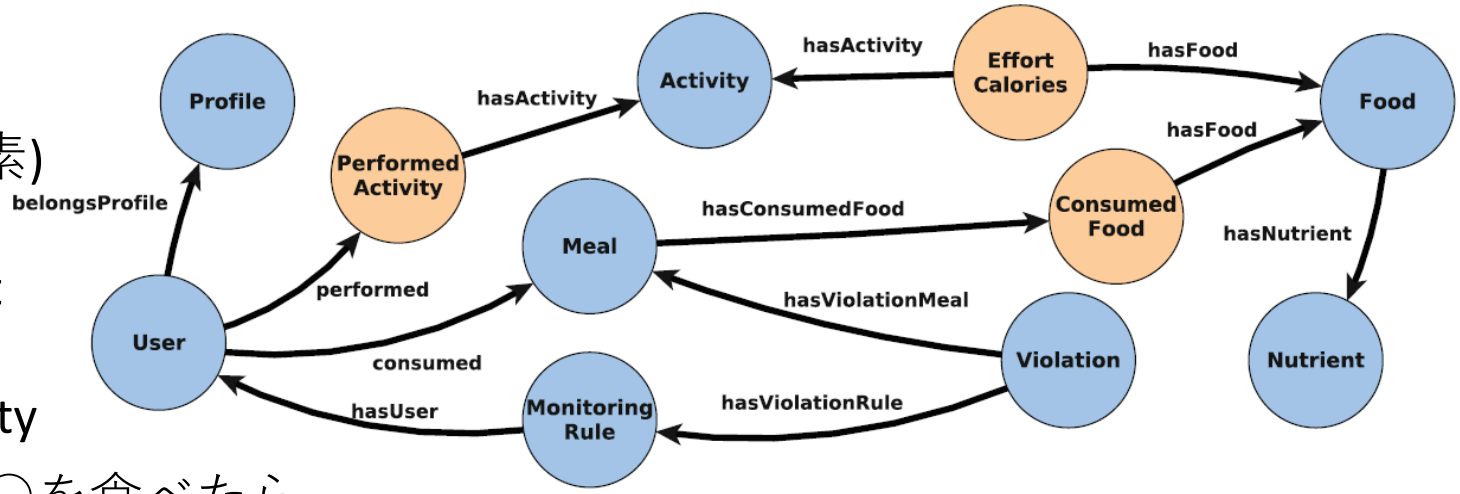


Fig. 1. Overview of the HeLiS ontology.

2.1.1. Ontologies & Data Integration

RESOURCES

Knowledge Integration for Disease Characterization: A Breast Cancer Example

Oshani Seneviratne, Sabbir Rashid, Shruthi Chari, Jim McCusker, Kristin Bennett, James Hendler and Deborah McGuinness

- **胸部癌のオントロジー．医師が患者のリスク，処置方法，モニタリング計画などを決定するのを支援するツールに利用．**
  https://cancer-staging-ontology.github.io/
- 複数リソース(薬品，遺伝子など)からデータを抽出・統合．
- OWLルール＋ドメインエキスパートによるSemantic Mapping

支援ツール

- Whyis Knowledge Graph Frameworkを利用
- Cancer Stagingガイドラインの複数版を比較可

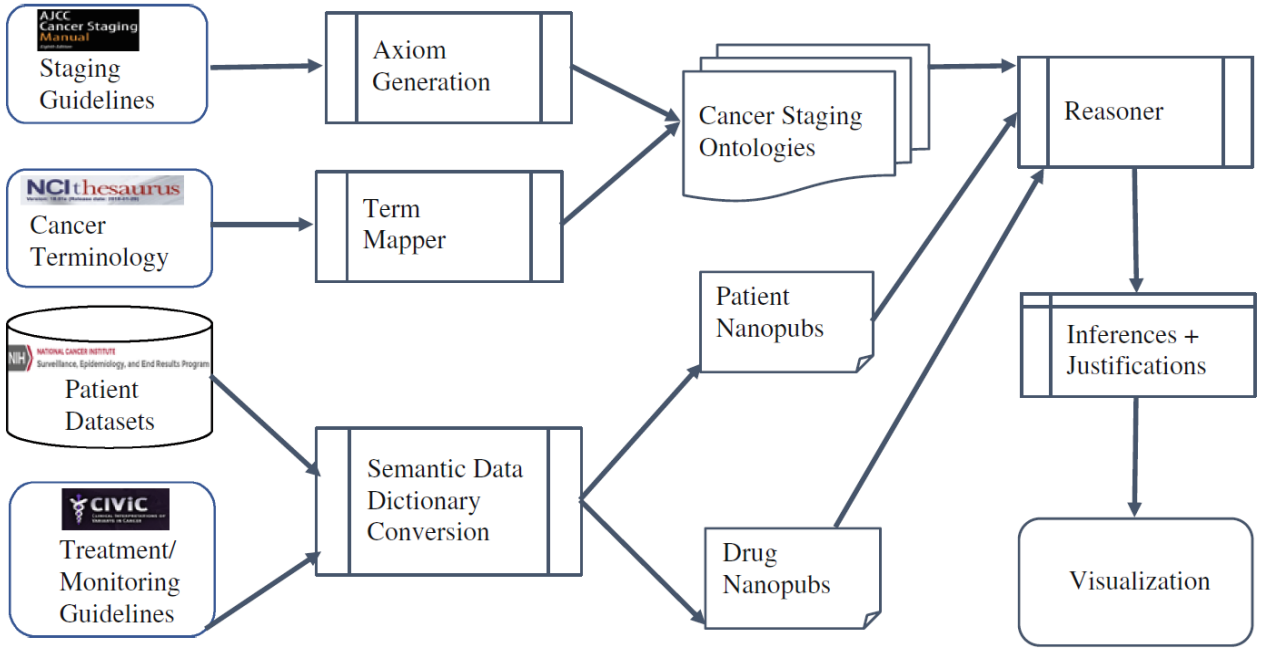


Fig. 1. Knowledge integration architecture for breast cancer characterization

3月11日(月) 14:30-15:30 ISWC2018サーベイ2

No.	担当者(所属)	担当セッション
7	加藤 文彦 (国立情報学研究所)	Oct10, Session 3.3 Scholarly Data Oct11, Session 1.3 Ontologies & Linked Data
8	山口 敦子 (ライフサイエンス統合データベースセンター)	Oct11, Session 1.2 Querying Oct12, Session 1.2 Querying
9	森田 武史 (慶應義塾大学)	Oct11, Session 2.2 Ontologies & Data Integration Oct12, Session 1.1 Ontologies & Data Integration
10	江上 周作 (電気通信大学)	Oct11, Session 2.4 Ontologies & Linked Data
11	Phuc Nguyen (NII)	Oct12, Session 2.1 Embeddings and Deep Learning
12	濱崎 雅弘 (産業技術総合研究所)	Oct12, Session 2.2 Information Extraction Oct12, Session 2.3 Scholarly Data
13	長野 伸一 (株式会社東芝)	Oct10, Session 2.4 Industry Oct10, Session 3.1 Embeddings and Deep Learning

Oct10, Session 3.3
Scholarly Data

Oct11, Session 1.3
Ontologies & Linked Data

7.加藤 文彦
(国立情報学研究所)

ISWC2018サーベイ発表

1.3.3 Scholarly Data
2.1.3 Ontologies & Linked Data

加藤文彦
国立情報学研究所オープンサイエンス基盤研究センター
2019-03-11

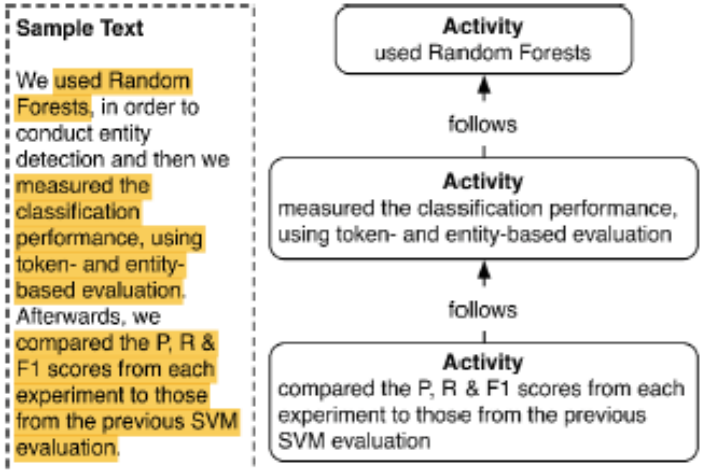
1.3.3 Scholarly Data

TSE-NER: An Iterative Approarch for Long-Tail Entity Extraction in Scientific Publications

- 科学出版物用の固有表現抽出・型付
  - 汎用手法ではWebKB: Organization, SimFusion+: なし
    - 理想はDataset,Software
  - 問題点:抽出対象がロングテール。科学用既存手法は人手によるお手製のヘビーなアノテーションが必要。
  - ロングテールな表現型付用分類器を訓練するために、少量のシーズを入力とした反復的アプローチを提案
- 10種カンファレンスシリーズの15994論文から"Dataset"と"Methods"を抽出。100個のシーズで
    - precision重視: 0.91, recall重視: 0.41, バランス: 0.77,0.30
  - わずか５個でもそこそこ性能がでる
- ドメイン例として生物医学4525論文についてProteinでも試験
  - 25個でprecision: 0.40, recall 0.28

Ontology Driven Extraction of Research Processes

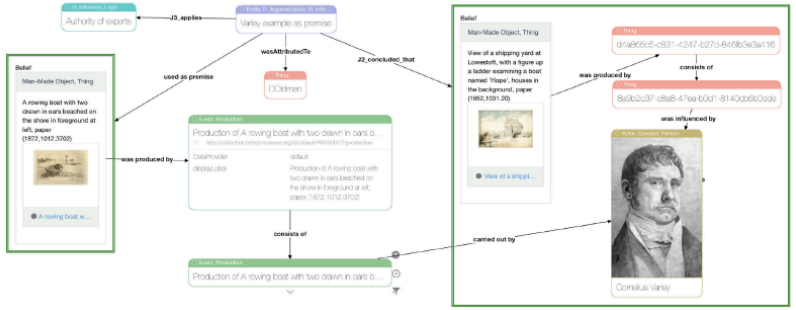
- 研究プロセス表現の2つの概念を出版物から自動抽出
  - 研究活動と継続的な活動間の順序関係をScholary Ontologyで表現
  - 複数の文にまたがって抽出する必要があり
  - Logistic Regression, SVM, Random Forestなど分類器を開発実験
  - 抽出した活動は他の関連情報(ORCID)にもリンクしてRDFに
- データセット: 50000 OA papers
  - 分野: DH, Bioinf, Med
  - Token-based評価: SVM-Pipeline (F1 0.88)
  - Entity-based評価: RF.PD.BS (F1 0.73)とSVM-Pipeline (F1 0.72)
  - 関係抽出評価:
    - Bioinf: RF.PD.B (F1 0.86)
    - Med: RF.PD.B (F1 0.92
    - DH: LR.WPD.B (F1 0.88)
    - 全体: RF.PD.B (F1 0.89)



IN-USE

Reshaping the Knowledge Graph by connecting researchers, data and practices in ResearchSpace

- ResearchSpace: 大英博物館が開発したオープンソース
  - https://www.researchspace.org/
  - データ共有、正式な議論への積極的な関与、セマンティック出版
  - 専門家による作成、表明、注釈、主張、検索、引用等をつけさせたい
  - CIDOC CRM, Open Annotation, IIIFなど
- 応用プロジェクト
  - 考古学: 文化遺産の幾何学的再構成と意味的な再統一
    - コレクション横断して3D画像の部品を知識グラフ化・アルゴリズム開発
  - 美術史: 葛飾北斎について内部・外部のリソースをリンク
  - 科学史: CorpusTracer, CIDOC CRM+FRBRoo



2.1.3 Ontologies & Linked Data

Detecting Erroneous Identity Links on the Web using Network Metrics

- LODクラウドから集めた558Mのowl:sameAsから誤りを検出
- identityネットワーク自体のコミュニティ構造を抽出
  - 単方向は1、双方向は2とした無向グラフ
  - 同等の集合に分割後、Louvainアルゴリズムでコミュニティ検出
  - コミュニティ内と外で重み付きリンク作成
  - コミュニティ内と外でリンクエラー度合い計算
- エラー例: バラック・オバマという人物、と大統領
- 73%のリンクがエラー度<=0.4で、その場合100%正しいリンク
- 1.2m のリンクが0.99-1のエラー度で、内88%が間違ったリンク
- https://sameas.cc (githubも含めてサイト落ちている?)

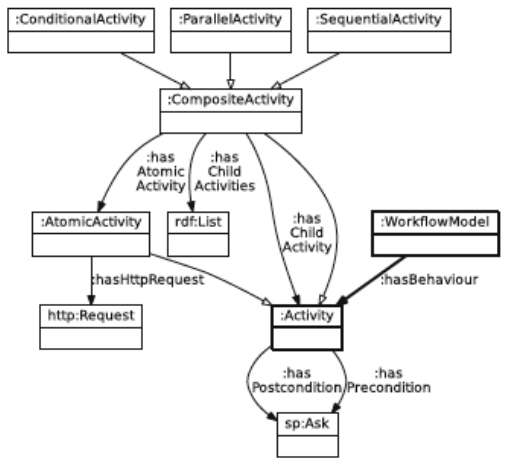
IN-USE

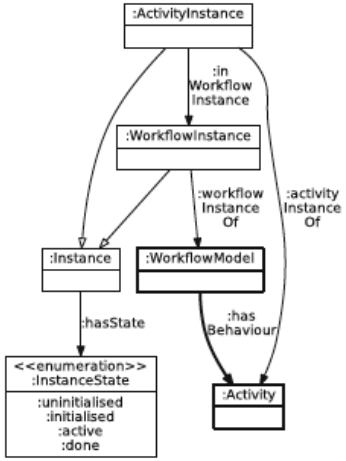
Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph

- WikidataのRDFとQuery Service
  - 400m statements, 45m entities, 100m queries/month
  - RDFダンプ 4.0b triples, 32GB in gzipped Turtle
- RDF: 内部表現はJSONなのでRDF用語彙を作成しマッピング
  - 複合データ値やqualifiersの問題
- WDQS: http://query.wikidata.org
  - Wikidataに特化したフォーム。マウスオーバーでの情報提示。結果の可視化など。
  - SPARQL1.1 + 独自のラベル検索。Federated query先のホワイトリスト
  - BlazeGraphを2つのデータセンターで6台マシン
    - 2018年1-3月のクエリ返答: Top 50%は40ms以下、95%は平均440ms、99%だと60s近くに
    - BlazeGraphはその拡張性と性能に満足している
  - 人間とロボットのクエリの分析: SERVICES langは人間は頻繁に使うが機械は使わない、逆にVALUESはロボットが使う、等

Specifying, Monitoring, and Executing Workflows in Linked Data Environments

- Read-Write Linked Data用ワークフロー表現オントロジー
  - RDLDの例: LDP, Solid, IoT/WoT
  - ASM4LD: RWLD用ルールベース言語
  - 操作的意味論の定義
- 応用: 仮想の航空機のコクピットデザイン
  - VR、フライトシミュレーション、センサー、ワークフローの統合サイバーフィジカルシステム
- ビル管理システム用のBrick ontologyに記述されているビルが表現できるか実験





AtomicActivity(a) ∧ hasHttpRequest(a,h) ∧ http:mthd(h,method)
∧ http:requestURI(h,u) ∧ ...→METHOD(u,...)
ldp:contains(http://example.org/ldpc,e)→GET(e)

IN-USE

Ontology-Based Recommendation of Editorial Products

- SpringerNatureのトピックと関連コミュニティの興味をマッチ
  - 従来はエディタによる手動だがCSのような変化の激しい分野は大変
- Smart Book Recommender (SBR)
  - 本、雑誌、論文集をカンファレンスに関連付
  - Computer Science Ontology
    - BIBO拡張、Klink2
    - 15K semantic topics, 70K relationships
  - Smart Topic API
    - 書誌情報投げると分類を返答
    - コサイン類似度
- ユーザ評価:7人のエディターと9人の研究者
  - 2つのカンファレンス選択=>20個の本と10個のカンファレンス推薦
  - Precision @10: 本 76.8%, 論文集 75.4%
  - 関連を正しいと判断した率: エディター72.9%, 研究者66.8%

Oct11, Session 1.2
Querying

Oct12, Session 1.2
Querying

8.山口 敦子
(ライフサイエンス統合データベースセンター)

ISWC 2018 サーベイ
Query2
Query3

ライフサイエンス統合データベースセンター
山口敦子

Research, Best Student Paper

Canonicalisation of Monotone SPARQL Queries

Jaime Salas and Aidan Hogan

- 背景と問題

SPARQLエンドポイントの利用の際は，しばしばそのパフォーマンス(応答速度)が問題になる．応答のキャッシュを行うとしても，SPARQLクエリの同一性を判定する必要がある(SPARQLクエリの同一性判定問題はNP完全).

- 方法

SPARQLクエリの正規化によって，意味的に同じクエリであれば同じクエリに正規化される手法を提案する.

- 結果

単調SPARQLクエリ(トリプルパタン，結合，和，射影のみからなるクエリ)の場合，同一の意味のクエリが同一クエリになる手法を提案した.

Comunica: A Modular SPARQL Query Engine for the Web

Ruben Taelman, Joachim Van Herwegen, Miel Vander Sande, and Ruben Verborgh

- 背景と問題

Linked Dataの検索インタフェイスとして様々(Linked Data documents, SPARQLエンドポイント，Triple Pattern Fragments)なものがあり，SPARQLエンドポイントだけでもストアごとの方言があり，一元的に利用しにくい.

- 方法

様々なLinked Dataの検索インタフェイスを扱える検索エンジンを作る.

- 結果

ComunicaというJavascriptベースのツールを作った.
GitHubでMITライセンスで公開中.
https://github.com/comunica/comunica

Querying Large Knowledge Graphs over Triple Pattern Fragments: An Empirical Study

Lars Heling(B), Maribel Acosta, Maria Maleshkova, and York Sure-Vetter

- 背景と問題

Triple Pattern Fragments(TPFs)が提案されて以降，そのパフォーマンスは主にSPARQLのクエリの応答速度について議論されてきた．TPFs自体の大規模なベンチマークと解析はされていない．

- 方法

TPFsの応答速度を6つの観点から調査した．1)トリプルパタンのタイプ, 2)応答のサイズ, 3)ページサイズ, 4)バックエンド, 5)複数のグラフを含むかどうか, 6)人工環境か実世界か

- 結果

1) バックエンドにかかわらず大きな影響がある, 2)やや関係ある, 3)バックエンドによる, 4)HDTがよい, 5)複数だと下がる, 6)実世界だと外因性のファクターにより下がる．

Journal

Effective searching of RDF knowledge graphs

Hiba Arnaout, Shady Elbassuoni

- 背景と問題

RDFには以下の4つの課題がある．1)データの不完全性(フリーテキストでは書けるがトリプルでは難しいなど), 2)クエリの柔軟性, 3)結果のランキング, 4)結果の多様性

- 方法

三つ組を五つ組にするフレームワークの提案

Table 2
An excerpt of DBpedia's extended RDF knowledge graph in the form of RDF triples.

Subject	Predicate	Object	Weight	Keyword:Weight
Annie_Hall	director	Woody_Allen	2032	oscar:1, comedy:251, york: 1449, romance: 13, ...
Annie_Hall	starring	Diane_Keaton	758	psychoanalysis:8, york:567, love:494, sex:8, jew:72, ...
Match_Point	writer	Woody_Allen	1866	thriller:121, greed:9, lust:32, money: 195, allen:2063, ...
Match_Point	starring	Emily_Mortimer	665	london:216, pregnancy:3, thriller:53, love:365, lust:20, ...
Match_Point	producer	Letty_Aronson	138	thriller:12, money:20, london:67, crime:14, sister:1, ...
Unstrung_Heroes	director	Diane_Keaton	517	1995:115, comic:30, memoir:9, journalist:47, ...

- 結果

DBpediaのデータを用いてフレームワークの有効性を確認

Journal

LargeRDFBench: A billion triples benchmark for SPARQL endpoint federation

Muhammad Saleem,
Ali Hasnain , Axel-Cyrille Ngonga Ngomo

- 背景と問題

Federated SPARQL検索のベンチマークテストはFedBenchが知られているが，ベンチマークテストとしてはデータのサイズが小さすぎる.

- 方法

13データセットで計1Bトリプルのベンチマークを作成した.このベンチマークを用いて，FedX, SPLENDID, ANAPSID, HiBISCuSの比較を行った.

- 結果

join-aware TPWSS(Triple Pattern Wise Source Selection)方式を採用したANAPSID, HiBISCuSが優れていた.

Journal

Impact Analysis of Data Placement Strategies on Query Efforts in Distributed RDF Stores

Daniel Janke, Steffen Staab, Matthias Thimm

- 背景と問題

一つのトリプルストアに入りきれない，多くのクエリをさばく必要があるなどの理由で，複数のノードに分散したストアに配置する必要が出てきた.

- 方法

下記の条件をできるだけ満たすような置き方を考える．1) トリプル数をバランスよく配置, 2) 別のクエリの結果は別のノードで並列に計算, 3)個々のクエリはできるだけ少ない数のノードで計算.

- 結果

既存の配置法: hash, hierarchical, minimal edge cutを上記の条件において比較し，従来言われていた仮説に反し，minimal edge cutよりhashがよい結果を出した.

Journal

OptiqueVQS: a Visual Query System over Ontologies for Industry

Ahmet Soylu, et al

- 背景と問題

データモデルをオントロジーで作り，アプリケーションデータをRDFで利用することが多くの大企業で行われてきた．その際，ドメイン専門家にオントロジーを利用して必要な情報を得るためのクエリを書かせることが重要である．しかし，ドメイン専門家はオントロジーやDBの知識がなく難しい．

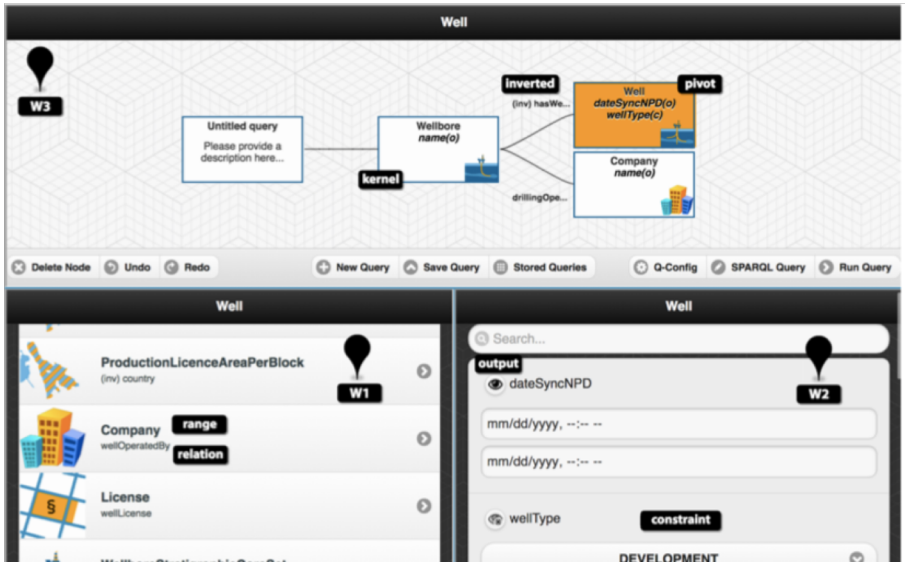
- 方法

視覚的なクエリ作成システムを提案する．

- 結果

OptiqueVQSというツールを開発した．

https://gitlab.com/ernesto.jimenez.ruiz/OptiqueVQS



Journal

Distributed Query Processing in the Presence of Blank Nodes

Audun Stolpe and Jonas Halvorsen

- 背景と問題

空白ノードを含むデータセットに対し，通常の分解手法でSPARQLクエリを分散処理すると，少ない結果を出すことがある．

- 方法

問題が起こることを形式意味論を用いて示す．問題の解決手法を集合意味論を用いて提案する．

- 結果

SPARQL1.1 意味論に形式パラメータを追加することで，空白ノードが特定できる手法が考えられる．しかし，この解決手法は計算量的に困難である．

Oct11, Session 2.2
Ontologies & Data Integration

Oct12, Session 1.1
Ontologies & Data Integration

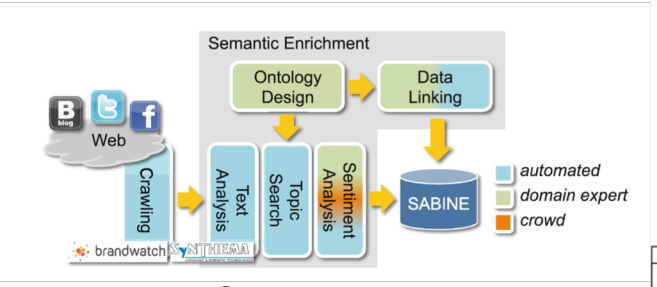
9.森田 武史
(慶應義塾大学)

SABINE: A Multi-Purpose Dataset of Semantically-Annotated Social Content

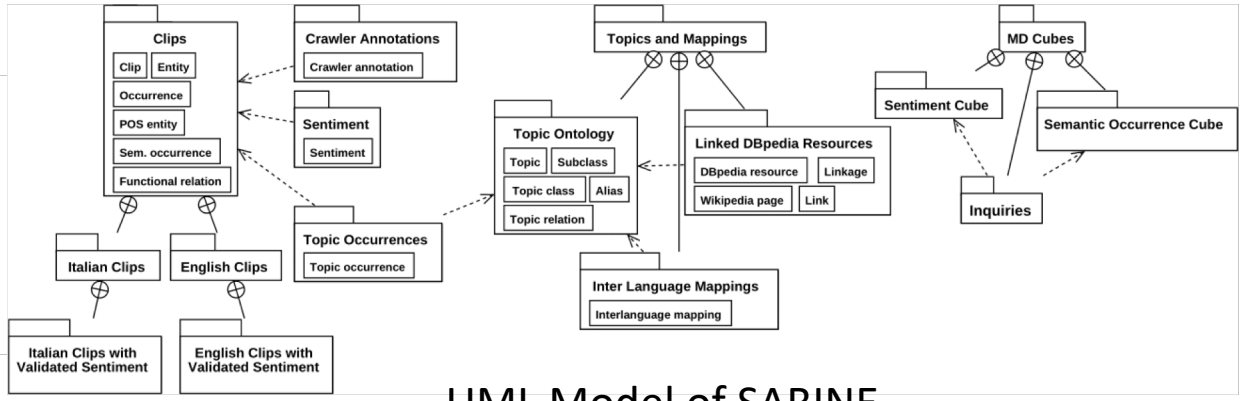
RESOURCES

Matteo Golfarelli, Stefano Rizzi, Enrico Gallinucci, Silvana Castano, Alfio Ferrara, Lorenzo Mosca, Cristian Vaccari and Stefano Montanelli (*University of Milan, Bologna, Roma Tre, London*)

- **背景**
  - SBI (**S**ocial **B**usiness **I**ntelligence): 経営意思決定のために，企業データとソーシャルデータを結び付ける分野
  - WebPolEuプロジェクト (http://www.webpoleu.net/)の成果の1つ
- **課題**
  - 入手可能な実世界のデータの不足
  - Ground truthを決定することが困難
- **提案・貢献**
  - **SABINE** (**S**oci**A**l **B**usiness **IN**telligence b**E**nchmark)
  - ヨーロッパの政治に関するSBIのためのデータセット
    - センチメントスコアとメタデータを付与したイタリア語と英語の600万クリップ
    - 400個のトピックを含むオントロジー
    - DBpediaとのマッピング
    - センチメントとsemantic occurrenceについて分析と集約をするための2つの多次元キューブ
    - クラウドソーシングと領域専門家によりデータを検証
  - CC BY-NC 4.0 licenseでデータセットを公開
    - http://big.csr.unibo.it/sabine



SBI process



UML Model of SABINE

http://iswc2018.semanticweb.org/sessions/sabine-a-multi-purpose-dataset-of-semantically-annotated-social-content/

SPgen: A Benchmark Generator for Spatial Link Discovery Tools

RESEARCH

Tzanina Saveta, Irini Fundulaki, Giorgos Flouris and Axel-Cyrille Ngonga Ngomo
(*FORTH, University of Paderborn*)

- **背景**
  - リンク発見システムの性能を評価するために多くのベンチマークが提案されてきたが，地理空間エンティティにリンクする問題を対象にしたベンチマークは少ない
  - ヨーロッパのH2020プロジェクトHOBBIT(https://project-hobbit.eu/)の一環
- **課題**
  - 空間リソースのトポロジーとそれらの間の関係を考慮して，地理空間リンク発見のためのベンチマークを開発すること
- **提案・貢献**
  - SPgen (**S**patial **B**enchmark **Gen**erator)
  - DE-9IM（Dimensionally Extended 9-Intersection Model）で提案されているLineStringとPolygonの位相関係を扱うリンク発見システムの性能をテストするためのベンチマークを生成
  - データセット
    - TomTom Data Generator
    - Spaten: Spatio-temporal and Textual Big Data Generator
  - 4つのリンク発見システム(AML, OntoIdea, RADON, Silk)をSPgenを用いて評価
  - https://github.com/hobbit-project/SpatialBenchmark

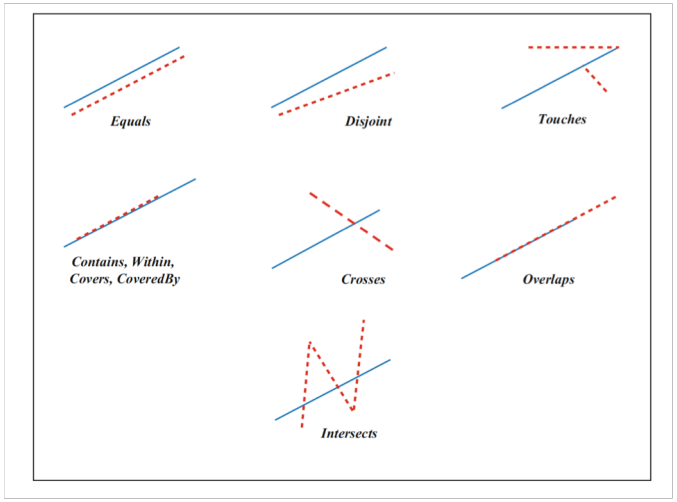


Fig.1. Examples of DE-9IM topological relations for LineStrings

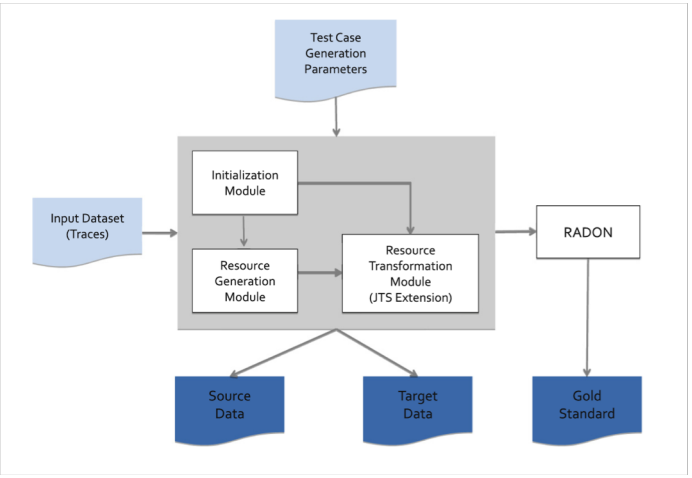


Fig.4. SPgen architecture

http://iswc2018.semanticweb.org/sessions/spgen-a-benchmark-generator-for-spatial-link-discovery-tools/

PrivOnto: a Semantic Framework for the Analysis of Privacy Policies

JOURNAL

Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B. Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, Norman Sadeh
(*Carnegie Mellon University, Fordham University School of Law, University of Michigan School of Information, University of Cincinnati*)

- **背景**
  - プライバシーポリシーは，長く複雑で多くのユーザは読んでいない
  - 半自動的にプライバシーポリシーに注釈を付けることが必要
- **課題**
  - プライバシー要件の特定と形式言語による表現
  - **システムからの関連情報抽出**
  - ユーザ嗜好に応じたプライバシーポリシーの分析と検証
- **提案・貢献**
  - **PrivOnto:** 注釈付きプライバシーポリシーを表現するためのセマンティックフレームワーク
    - ユーザまた法律専門家にとって重要な問題を体系化したプライバシーポリシーオントロジー
    - プライバシーポリシー注釈支援ツール
    - 注釈付きプライバシーポリシー検索ツール
    - プライバシーポリシーオントロジーを用いた推論
  - **UPP**(Usable Privacy Policy)により公開されている23,000を超える注釈付きプライバシーポリシーコーパス(115セット)により検証
    - https://explore.usableprivacy.org/
  - 約60個のSPARQLクエリを用意し，ユーザが関心のあるプライバシーに関する質問に回答
  - 大規模なプライバシーポリシー分析

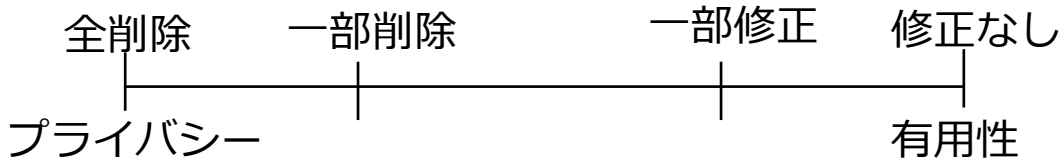
http://iswc2018.semanticweb.org/sessions/privonto-a-semantic-framework-for-the-analysis-of-privacy-policies/

Query-based Linked Data Anonymization

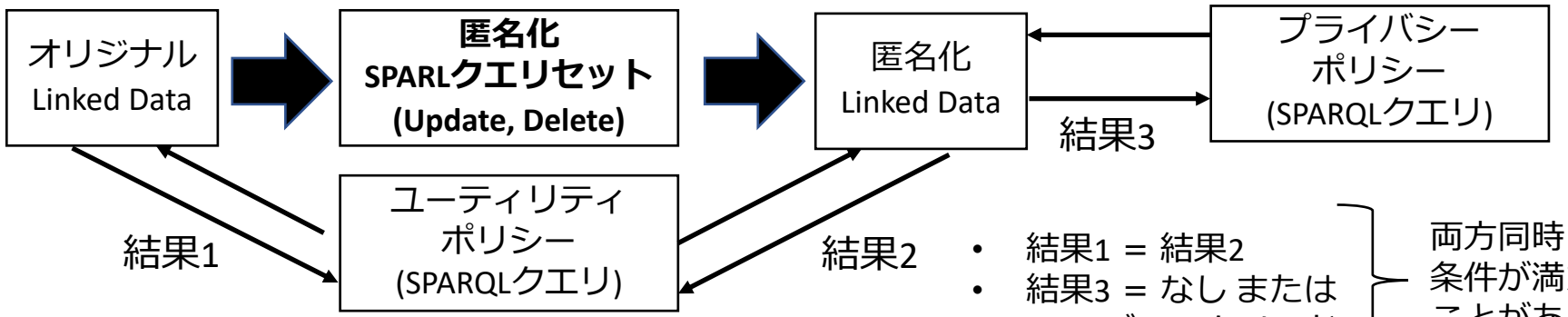
RESEARCH

Rémy Delanaux, Angela Bonifati, Marie-Christine Rousset and Romuald Thion
(*Universite Lyon 1, Universite Grenoble Alpes, Institut Universitaire de France*)

- **背景**
  - 多くの組織や政府はLODを公開しているが、ID開示のリスクは完全には理解されていない
- **課題**
  - プライバシーと有用性の両方のバランスを取りながら匿名化操作を行うこと



- **提案・貢献**
  - プライバシー保護されたLinked Data公開のための宣言的フレームワーク
    - https://github.com/RdNetwork/Declarative-LOD-Anonymizer



- 結果1 = 結果2
- 結果3 = なし または ブランクノード

両方同時に条件が満たせないことがある

http://iswc2018.semanticweb.org/sessions/query-based-linked-data-anonymization/

Practical Ontology Pattern Instantiation, Discovery, and Maintenance with Reasonable Ontology Templates

RESEARCH

ベストペーパー候補

Martin G. Skjæveland, Daniel P. Lupp, Leif Harald Karlsen and Henrik Forssell (*University of Oslo*)

- **背景**
  - 維持可能な高品質の大規模オントロジー構築は困難
- **課題**
  - オントロジーの構築と保守のためにツールにより支援されるベストプラクティスの欠如
  - オントロジー構築における3つのフェーズ：
    1. 対象領域の理解，2. 領域における抽象概念の特定，**3. 形式言語 (例: DL)による抽象化の定式化**
  - 例：ピザオントロジーは22種類のピザが，同一公理パターンに従って定義されているが，この公理パターンは明示的に表現されていない
    → パターンの検索，パターンの一貫性検証，オントロジーの理解，パターンの更新などが困難
- **提案・貢献**
  - **OTTR** (Reasonable Ontology Templates) (https://ottr.xyz/)
  - パラメータ化されたオントロジーの形式でオントロジーのモデリングパターンを表現するための言語
  - ピザオントロジーのOTTR
    - http://sws.ifi.uio.no/ottrs/info/?tpl=http://draft.ottr.xyz/pizza/NamedPizza

```
NamedPizza(
    ?Name : 1 class,
    ?Country : ? individual,
    ?Toppings : + class)
::
    SubClassOf(?Name, :NamedPizza),
    SubObjectHasValue(?Name, :hasCountryOfOrigin, ?Country),
    SubObjectAllValuesFrom(?Name, :hasTopping, _:b1),
    ObjectUnionOf(_:b1, ?Toppings),
    x | SubObjectSomeValuesFrom(?Name, :hasTopping, ?Toppings) .

NamedPizza(:Margherita, :Italy, ⟨:Tomato, :Mozzarella⟩)
NamedPizza(:Grandiosa, none, ⟨:Tomato, :Jarlsberg, :Ham, :SweetPepper⟩)
```

(a) stOTTR serialisation and instances

```
:pizza  rdfs:subClassOf  p:NamedPizza ,
     [ a                    owl:Restriction ;
       owl:onProperty       p:hasTopping ;
       owl:someValuesFrom   :toppings ] ,
     [ a                    owl:Restriction ;
       owl:onProperty       p:hasTopping
       owl:allValuesFrom    [
           a              owl:Class ;
           owl:unionOf    ( :toppings ) ] ] ,
     [ a                  owl:Restriction ;
       owl:onProperty     p:hasCountryOfOrigin ;
       owl:hasValue       :country ] .
p:hasTopping          a  owl:ObjectProperty .
p:hasCountryOfOrigin  a  owl:ObjectProperty .
:toppings             a  owl:Class .
```

(c) Expanded RDF graph

http://iswc2018.semanticweb.org/sessions/practical-ontology-pattern-instantiation-discovery-and-maintenance-with-reasonable-ontology-templates/

Mapping Diverse Data to RDF in Practice

RESEARCH

Alexandros Chortaras and Giorgos Stamou (*National Technical University of Athens*)

- **背景**
  - R2RML: 関係データベースをRDFにマッピングする言語
    - https://www.w3.org/TR/r2rml/
  - R2RMLの拡張としてxR2RMLやRMLが提案されている
    - NoSQL, CSV, XML, JSON, HTMLからRDFへの変換
- **課題**
  - 既存のRDFマッピング言語は，データソースのデータ構造が維持されることを仮定（特にRMLの問題点）
- **提案・貢献**
  - D2RML (**D**ata-to-**R**DF **M**apping **L**anguage)
  - R2RMLとRMLを拡張し，多様なデータソースやWebサービスからデータを選択して統合し，RDFデータストアを生成
- **参考**
  - 本論文はRMLの著者らの最近の論文を引用していない
  - 本論文で挙げているRMLの問題点は全て最近の論文で対応
    - http://rml.io/blog/2018/d2rml-comparison.html
  - https://ci.mines-stetienne.fr/sparql-generate/

A Novel Approach and Practical Algorithms for Ontology Integration

RESEARCH

Giorgos Stoilos, David Geleta, Jetendr Shamdasani and Mohammad Khodadadi(*Babylon Health*)

- **背景**
  - Babylon Healthでは，オントロジーマッピング手法を用いて，既存の医学データ(SNOMED, UMLSなど)を統合して大規模な医学用KBを構築する試みがなされている
- **課題**
  - 新規に医学データをKBに統合する場合に，既存のサービスに影響を与えてはいけない
  - KBにはラベルの重複が多いエンティティが多数含まれてはいけない
    - 診断目的のKBを作成するためのテキストのアノテーションや医師によるクラス選択が複雑になるため
- **提案・貢献**
  - アプリケーションで利用されているオントロジーに，初期の構造や公理をなるべく保持しつつ，新しいデータを繰り返し統合して拡張するためのフレームワーク
  - オントロジーを統合する際の不整合を解消するために，既存のアプローチはマッピングを削除するが，ラベルの重複が発生する可能性があるため，本研究では公理の削除を行う
  - Babylon Healthでは，本フレームワークを利用して多くの医学オントロジーを統合し，ヘルスケアサービスを提供

http://iswc2018.semanticweb.org/sessions/a-novel-approach-and-practical-algorithms-for-ontology-integration/

Pragmatic Ontology Evolution: Reconciling User Requirements and Application Performance

Francesco Osborne and Enrico Motta (*The Open University*)

RESEARCH

- **背景**
  - データ統合，推論，分類，推薦，クラスタ分析などを行うために，オントロジーを用いて大規模な商品カタログを記述する組織が増えている
  - 領域の変化，ユーザやアプリケーションからの新しい要求に伴ってオントロジーは定期的に保守する必要がある（ontology evolution）
- **課題**
  - Ontology evolutionの3つの課題：保守の必要性を検出，変更候補の特定，変更の検証と評価
  - 新版のオントロジーがアプリケーションの性能に与える影響が評価されていない
  - 既存手法は単一概念の関連性を重視し，概念の組み合わせによる相乗効果を無視している
- **提案・貢献**
  - **POE** (**P**ragmatic **O**ntology **E**volution)フレームワーク
  - ユーザーの要求とそれが計算タスク（インスタンスタギング，類似度計算，推薦，データクラスタリングなど）に与える影響を考慮して，ontology evolutionのための概念選択を行う
  - Springer NatureのPMC (Product Market Codes) タクソノミーを対象に評価

http://iswc2018.semanticweb.org/sessions/pragmatic-ontology-evolution-reconciling-user-requirements-and-application-performance/

Oct11, Session 2.4

Ontologies & Linked Data

10.江上 周作
(電気通信大学)

ISWC2018サーベイ発表
- Ontologies & Linked Data -

江上周作
電気通信大学大学院情報理工学研究科

Ontologies & Linked Data (1/4)

A Framework to Build Games with a Purpose for Linked Data Refinement

Best Resource Paper

– Gloria Re Calegari, Andrea Fiano, and Irene Celino

■ 背景

- Linked Dataおよびナレッジグラフの増加により，欠損補完や誤り訂正の需要が高まっている
- これまでにいくつかの機械学習やクラウドソーシングによるナレッジグラフリファインメント手法が提案されている

■ 貢献

- Game with a purpose (GWAP)形式のクラウドソーシングでLinked Dataをリファインメントするための、オープンソースかつ再利用可能なフレームワークを提供（GitHub公開）
- 手作業でラベル付けしてgold standardを集める作業を大幅に省力化できる
- 使用ケースを紹介し，一般化のレベルと再利用の可能性，拡張性について説明

Ontologies & Linked Data (2/4)

DistLODStats: Distributed Computation of RDF Dataset Statistics

– Gezim Sejdiu, Ivan Ermilov, Jens Lehmann, and Mohamed Nadjib Mami

- 背景
  - Semantic Webは成長し，現在では10,000を超えるデータセットが利用可能になっているが，データセットの内部構造やカバレッジに関する統計情報などの情報が無いと活用が難しい
  - これらの情報を提供する既存ツールがあるが、サイズが大きくなると使用困難
- 貢献
  - 大規模RDFデータセットの統計計算のためのアルゴリズムを提案し，これを大規模分散，インメモリ計算するためのフレームワーク「DistLODStats」をApache Sparkを使って実装
  - 評価を行い，集中型アプローチに対する優位性を実証
  - SANSAフレームワークに統合．4つのシナリオ紹介．

Ontologies & Linked Data (3/4)

Browsing Linked Data Catalogs with LODAtlas

– Emmanuel Pietriga, Hande Gözükan, Caroline Appert,Marie Destandau, Sejla Cebiric, Francois Goasdoué, and Ioana Manolescu

■ 背景

- LODクラウドは急速に進化したが，エンドユーザが関連するデータセットを見つけることが難しくなっている
- セマンティックweb検索エンジンやデータカタログなどは役立つが，VoIDがついていなかったり，サービス終了していたり

■ 貢献

- データセットの中身やメタデータに対して，キーワードやファセット検索などを通して検索し，結果を視覚的に表示する「LODAtlas」を開発
- Linked Dataカタログをブラウジングするための単一的なWeb UIを提供し，データセットの断片や，カタログメタデータ，制約，相互関係などを可視化できる

Ontologies & Linked Data (4/4)

VoCaLS: Vocabulary and Catalog of Linked Streams

– Riccardo Tommasini, Yehia Abo Sedira, Daniele Dell'Aglio, Marco Balduini, Muhammad Intizar Ali, Danh Le Phuoc, Emanuele Della Valle, and Jean-Paul Calbimonte

■ 背景

- センサデータ分析，ソーシャルネットワーク，IoTなど，ストリームへの関心が高まり，Semantic WebコミュニティではStream Reasoningが進展している
- RDFストリームデータのためのRESTfulサービスIF，プロビジョニング，公開戦略など様々なアプローチがあるが，アクセスポイントおよび方法に関する共通のメタデータやオントロジーが定義されていない

■ 貢献

- 近年のチャレンジと課題を考慮した要件を提示
- ストリームおよびストリームサービスのリンキング，カタログ化のための語彙定義
- ストリームを発見可能，アクセス可能，再利用可能にするカタログ作成のためのロードマップを描く

Oct12, Session 2.1

Embeddings and Deep Learning

11. Phuc Nguyen
(NII)

SWO 47
Paper Review

ISWC 2018
Section: Embedding and Deep Learning (October 12, 2018)

Phuc Nguyen* (国立情報学研究所)

* https://sites.google.com/view/nguyentriphuc/

Embedding and Deep learning

Embedding and Deep Learning

1. Towards Encoding Time in Text-Based Entity Embeddings

   Federico Bianchi, Matteo Palmonari and Debora Nozza

2. Rule Learning from Knowledge Graphs Guided by Embedding Models

   Vinh Thinh Ho, Daria Stepanova, Mohamed H. Gad-Elrab, Evgeny Kharlamov and Gerhard Weikum

3. A Novel Ensemble Method for Named Entity Recognition and Disambiguation based on Neural Network

   Lorenzo Canale, Pasquale Lisena and Raphaël Troncy

Towards Encoding Time in Text-Based Entity Embeddings

Federico Bianchi, Matteo Palmonari and Debora Nozza

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Goal:

Encoding time (year) with entity embeddings

Novelty:

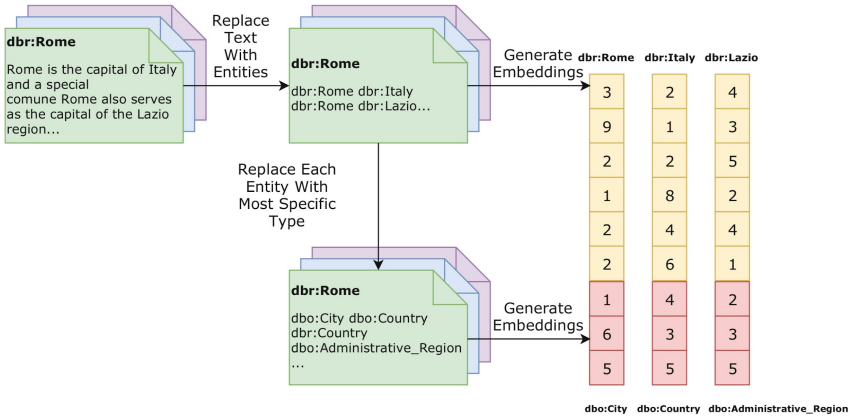
Incorporating time information

→ Improve the performance of other NLP tasks

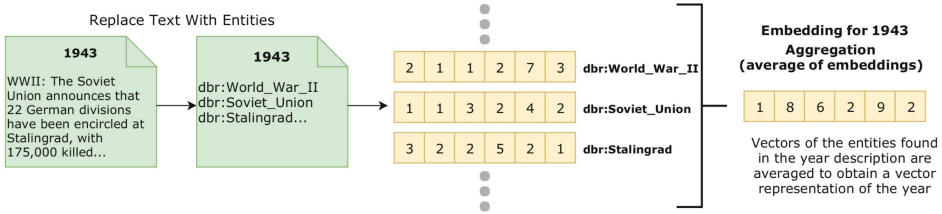
Techniques:

1. Entity embedding:
   Skip-gram: entity + type of entities
2. Year embedding:
   Average vector of all entities in this years

Entity embedding



Year embedding



Towards Encoding Time in Text-Based Entity Embeddings

Federico Bianchi, Matteo Palmonari and Debora Nozza

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

My thoughts:

- Entity embedding:

+ One entity can have multiple types. (Wikidata)

  Type(Italy) = {1.sovereign_state, **2.country,** 3.social_state, 4.unitary_state, 5.state, 6.republic}

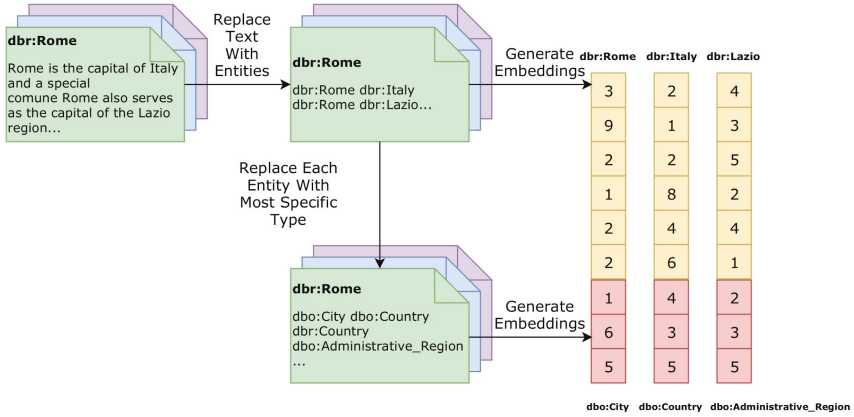
- Year embedding:

+ First contribution on encoding time.
+ Time concept: Complex
  - hour, day, month, granularity, semester, season

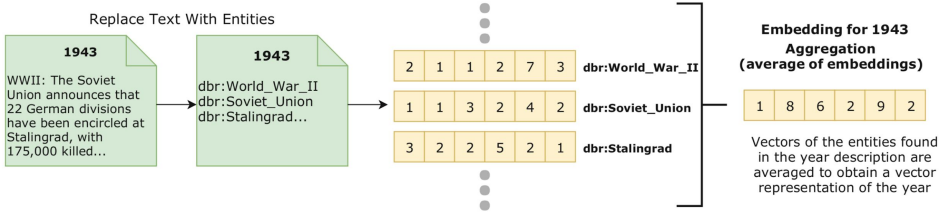
- Jointly embedding:

text corpus + knowledge graph

Entity embedding



Year embedding



Rule Learning from Knowledge Graphs Guided by Embedding Models

Vinh Thinh Ho, Daria Stepanova, Mohamed H. Gad-Elrab, Evgeny Kharlamov and Gerhard Weikum

Goal:

Rule learning in uncompleted graph

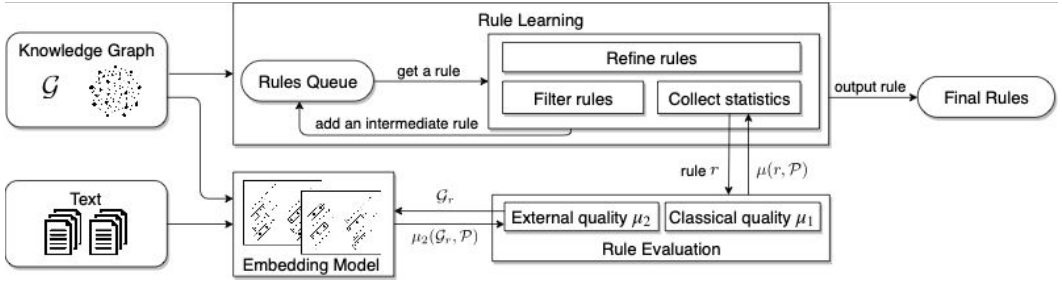
Novelty:

- Can learn nonmonotonic rules
- Rule mining + retrained embedding

Techniques:

1. Rule Learning (AMIE+)
2. Remove unpromising rule with Knowledge Base Embedding (TransE, HolE, SSP)

Framework



Example of RuLES:

r^1: $nationality(X, Y) \leftarrow graduated_from(X, Z), in_country(Z, Y), \textbf{not}\ research_uni(Z)$
r^2: $scriptwriter_of(X, Y) \leftarrow preceded_by(X, Z), scriptwriter_of(Z, Y), \textbf{not}\ tv_series(Z)$
r^3: $noble_family(X, Y) \leftarrow spouse(X, Z), noble_family(Z, Y), \textbf{not}\ chinese_dynasties(Y)$

Rule Learning from Knowledge Graphs Guided by Embedding Models

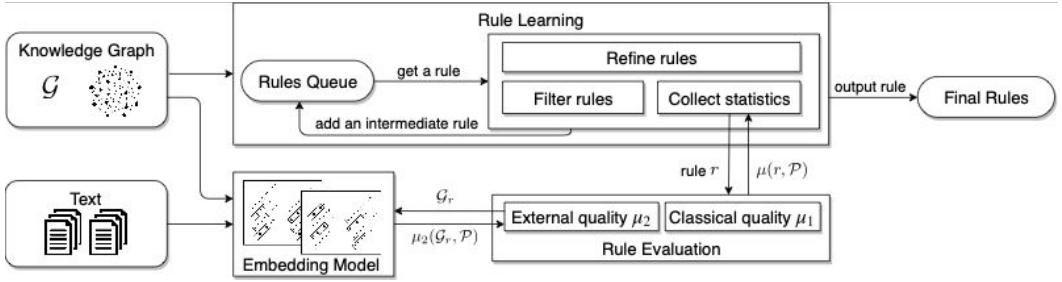
Vinh Thinh Ho, Daria Stepanova, Mohamed H. Gad-Elrab, Evgeny Kharlamov and Gerhard Weikum

My thought:

- This study:
  Enhance rule mining with embedding
- Invese study:
  Enhance embedding with rules

→ Rules + Embedding:
  Good for both
  - rule mining and
  - knowledge base embedding

Framework



Example of RuLES:

r^1: $nationality(X, Y) \leftarrow graduated_from(X, Z), in_country(Z, Y), \textbf{not}\ research_uni(Z)$
r^2: $scriptwriter_of(X, Y) \leftarrow preceded_by(X, Z), scriptwriter_of(Z, Y), \textbf{not}\ tv_series(Z)$
r^3: $noble_family(X, Y) \leftarrow spouse(X, Z), noble_family(Z, Y), \textbf{not}\ chinese_dynasties(Y)$

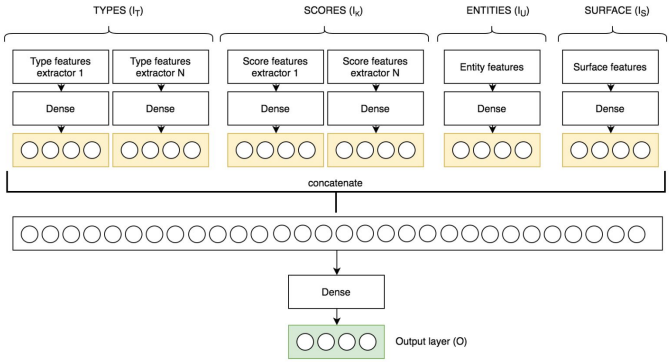
A Novel Ensemble Method for Named Entity Recognition and Disambiguation based on Neural Network

Lorenzo Canale, Pasquale Lisena and Raphaël Troncy

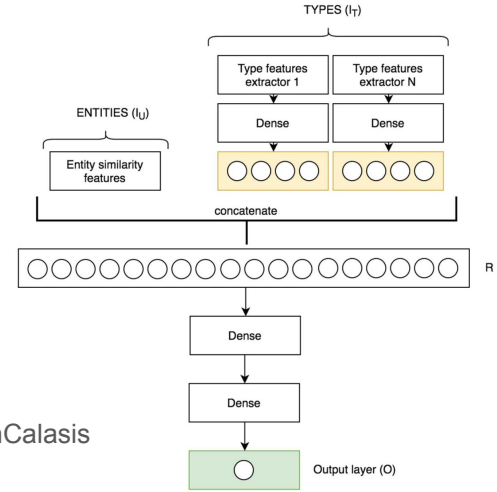
Goal:
Build two ensemble system from Open API* for two tasks:
- Name entity recognition (NER)
- Name entity disambiguation (NED)

Novelty:
- Combination systems from diffrent services (Industry)

Techniques:
1. Use open API for NER and NED
2. Use fastText (facebook) to convert the results of open API to vectors
3. Concate all vectors and train with simple neural network (1, or 2 layer)



NER Framework



NED Framework

* Ahchemy, Dandelion, DbSpotlight, TextRazor, Babelfy, MeaningCloudm, ADEL, OpenCalasis

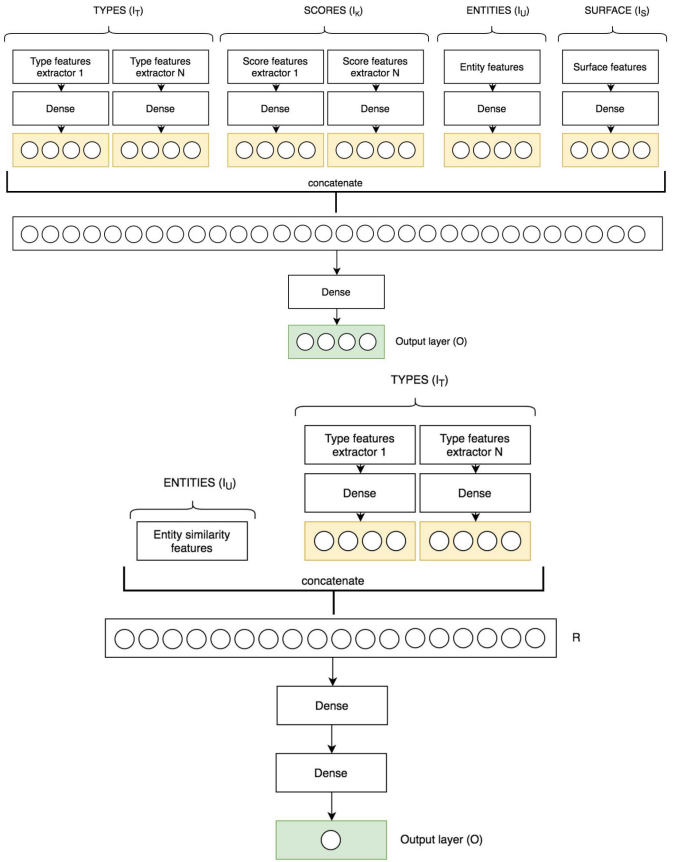
My thought:

- Potential:

Prasitce or industry

- Limitation:

+ Strongly depend on
  Open API services
  Word Embedding model
+ Cannot use when API can not access
+ Not opimamal for efficeny



NER Framework

NED Framework

Oct12, Session 2.2
Information Extraction

Oct12, Session 2.3
Scholarly Data

12.濱崎 雅弘
(産業技術総合研究所)

ISWC2018サーベイ

3.2.2 Information Extraction
3.2.3 Scholarly Data

濱崎雅弘

産業技術総合研究所

SIGSWO (2019.03)

3.2.2 Information Extraction

- Chair: Marieke van Erp

- [R] Structured Event Entity Resolution in Humanitarian Domains
  - Mayank Kejriwal, Jing Peng, Haotian Zhang and Pedro Szekely

- [J] A Framework for Real-time Semantic Social Media Analysis
  - Diana Maynard, Ian Roberts, Mark Greenwood, Dominic Rout, Kalina Bontcheva

- [R] That‘s Interesting, Tell Me More! Finding Descriptive Support Passages for Explaining Knowledge Graph Relationships ★
  - Sumit Bhatia, Purusharth Dwivedi and Avneet Kaur

Structured Event Entity Resolution in Humanitarian Domains

- **Background and Purpose**
  - 人道支援や災害救援（HADR）分野ではNamed EntityよりEventが重要
  - イベント同定技術Structured Event Entity Resolution (SEER)が必要
  - 本研究では、イベントクラスタリング手法AugSEERを提案
- **Contributions/Characteristics**
  - テキストデータ、グラフデータ、専門家による教師データを活用した手法
  - 5つのHADRデータセットを用いた評価実験、提案手法がBaselineを上回る
  - テキストデータだけでは良い結果が出なかった
    - Knowledge Graphの重要性

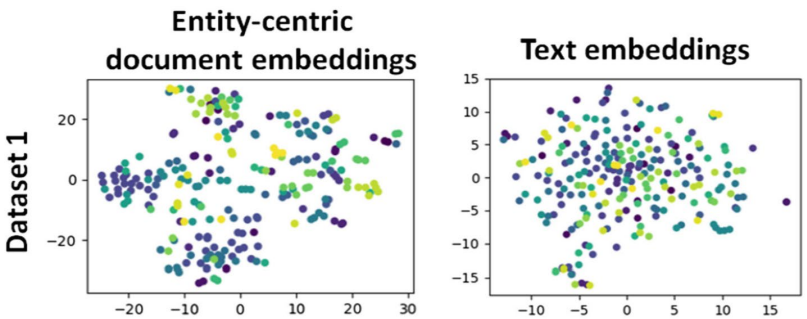


Fig. 3. t-SNE visualizations of all datasets using node/text embeddings

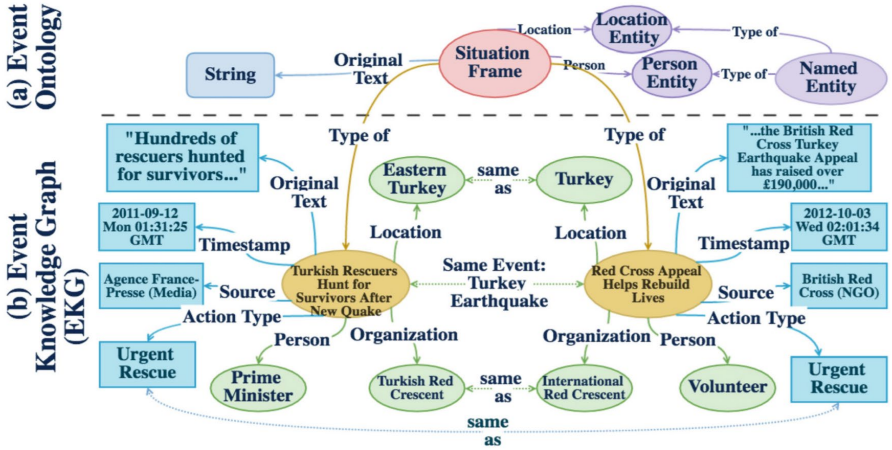


Fig. 2. A schematic illustrating the key representational details of the event ontology and event knowledge graph (EKG) for supporting solutions to the SEER problem.

That's Interesting, Tell Me More! Finding Descriptive Support Passages for Explaining Knowledge Graph Relationships

- **Background and Purpose**
  - Knowledge Graph で示された知識を説明できることは重要
    - ハイリスクな分野での意思決定支援や自動システムのための知識ベースでは致命的
  - 本研究では、関係（Triple）の説明に適したPassageを検索する手法を提案
- **Contributions/Characteristics**
  - 句、文書、コーパスそれぞれについてクエリとの関連性を求めて推定に利用
  - Wikidata/Wikipediaを用いて評価実験
    - クエリにWikidataから50個の関係を、コーパスにWkipediaを用いる
    - 提案手法のPrecisionは0.72、Baseline（Indri, 2004）との比較でMRRが3倍
- **Future Work**
  - Entity間の関係だけでなく、EntityとData valueの関係も扱う
  - Relation間の関係も利用する
    - <X, spouseOf, Y >
    - 「X is not wife of Y」

$$P(w|p) = P(w|\Theta_{MM}) \tag{4}$$

$$= \lambda_1 \underbrace{P(w|\Theta_p)}_{\text{passage-level evidence}} + \lambda_2 \underbrace{P(w|\Theta_d)}_{\text{document-level evidence}} + \lambda_3 \underbrace{P(w|\Theta_c)}_{\text{collection-level evidence}} \tag{5}$$

$$\text{Passage Evidence: } P(w|\theta_p) = \frac{count(w,p)+1}{|p|+|V|} \tag{6}$$

$$\text{Document Evidence: } P(w|\theta_d) = \frac{count(w,d)+1}{|d|+|V|} \tag{7}$$

$$\text{Collection Evidence: } P(w|\theta_c) = \frac{count(w,c)}{|C|} \tag{8}$$

3.2.3 Scholarly Data

- Chair: Anna Lisa Gentile

- [J] The Rijksmuseum Collection as Linked Data
  - Chris Dijkshoorn, Lizzy Jongma, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Wesley ter Weele, Jan Wielemaker

- [R] The Computer Science Ontology:
  A Large-Scale Taxonomy of Research Areas ★
  - Angelo Antonio Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne and Enrico Motta

- [R] The SPAR Ontologies
  - Silvio Peroni and David Shotton

The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas

- **Background and Purpose**
  - 研究分野オントロジーは、研究の探索・俯瞰のために重要
  - 多くのTaxonomyは数も少なく更新も遅い（ACMカテゴリは2000トピック、最新版が2012年作）
  - 本研究では、巨大なコンピュータ科学オントロジー（Computer Science Ontology）を自動生成
- **Contributions/Characteristics**
  - 1600万件の論文データセットにKlink-2を適用
  - 26,000個のトピック、226,000個の意味的関係
  - CSO Portal を公開
- **Future work**
  - CSに限らず全分野を網羅したオントロジー構築

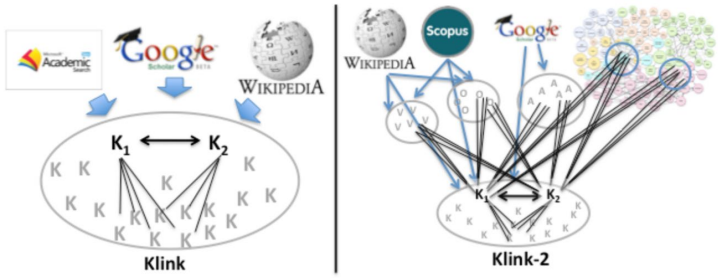


Figure 1. Relationships used for inferring the topic ontology in Klink and Klink-2.

[Osborne, ISWC2015]

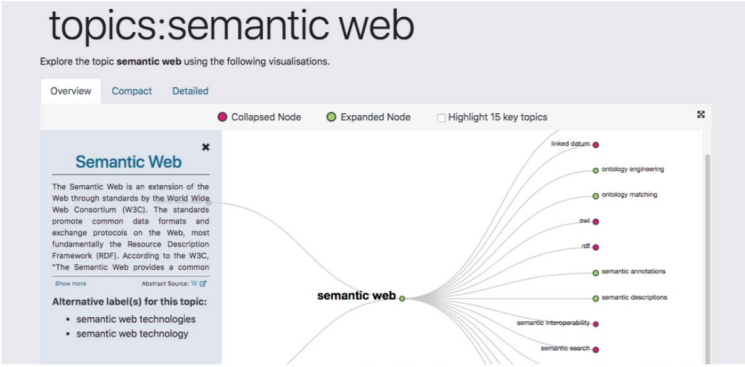


Figure 3. Screenshot of the resource page related to the topic *"semantic web"* (Overview).

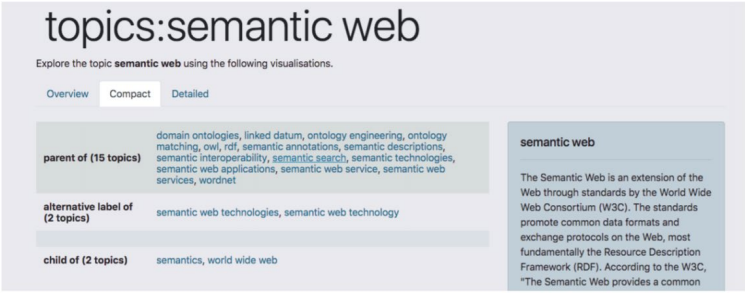


Figure 4. Screenshot of the resource page related to the topic *"semantic web"* (Compact).

The SPAR Ontologies

- **Background and Purpose**
  - 学術出版はこの10年で変革を迎え、出版・共有方法が劇的に変化
  - 学術情報の機械可読化は、論文査読者や読者にも多くの恩恵をもたらす
  - 本稿では、当該分野のためのSPAR（Semantic Publishing and Referencing）オントロジーを紹介
- **Contributions/Characteristics**
  - 2015年12月にWebサイト公開、
    現在までに100万アクセス（USが約8割）
  - SPARの開発ガイドライン
    - Enabling Adoption
    - Enabling Reuse
    - Enabling Application in Different Context
    - Enabling Interoperability
    - Enabling Human Understandability
- **Future work**
  - 著者や出版社がSPAR-basedメタデータを
    半自動で付与できるツールの開発

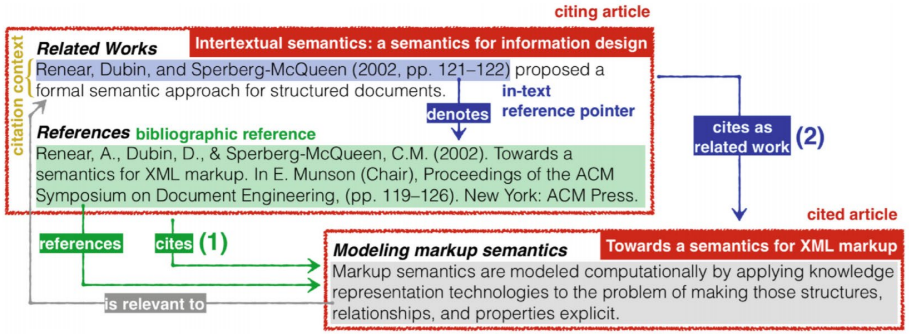


Fig. 3. The main components that allow a clear description of a citation and of all the related information.

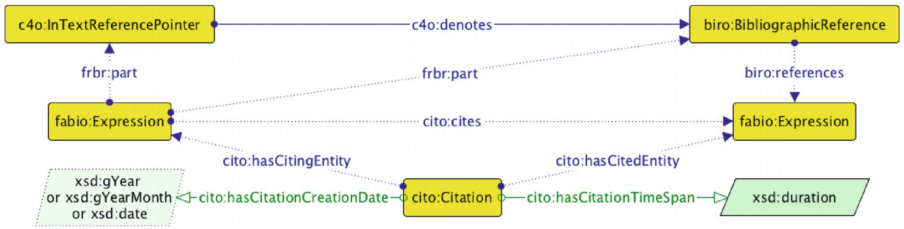


Fig. 4. A Graffoo diagram illustrating some of the ontological entities defined in CiTO, BiRO, and C4O.

Oct10, Session 2.4
Industry

Oct10, Session 3.1
Embeddings and Deep Learning

13. 長野 伸一
(株式会社東芝)

人工知能学会
第47回セマンティックウェブとオントロジー研究会
企画セッション「ISWC2018サーベイ」
担当セッション論文紹介資料

2019/03/11 東芝 長野

国際会議International Semantic Web Conference 2018
の論文概要をまとめたもの

1.2.4 Industry

Semantic Material Master Data Management at Aibel

University of Oslo, DLV GL*1, Aibel*2
*1: 第三者認証機関, Oil&Gasリスクマネジメント, 船級協会等のサービス企業
*2: Oil&Gas産業向けEPCサービス企業

■概要

- Oil&Gasプラットフォーム向けMMDオントロジーの構築・活用についての紹介

■課題

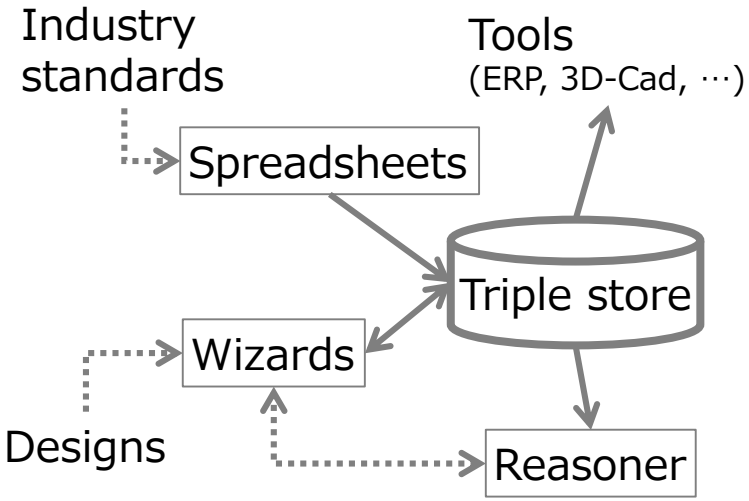
- 業務プロセスの生産性と品質の向上
  (従来は人手/文書ベースでムダが多かった)

■解決手段

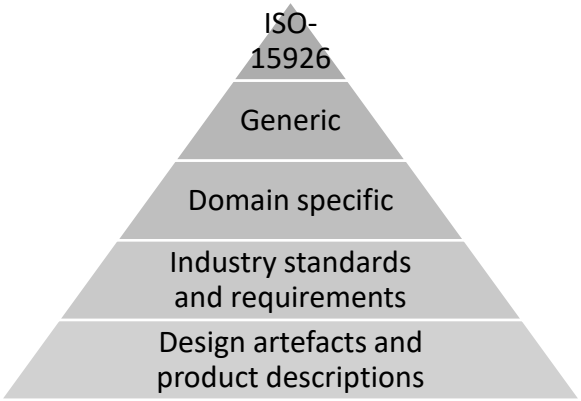
- 2012年からMMDオントロジーの構築を開始
- 仕様書/設計書の電子化・構造化を推進

■実績

- 全EPCプロジェクトで実運用中
- 設計図の誤り率が15%から0.5%へ減少
- 設計データの重複を削減，バルク注文のコスト削減(あるPJで1億€の削減効果)



システム構成*3



MMDオントロジー*3

*3: 両図とも論文の内容をもとに発表者が作成

From Data Search to Data Showcasing: The Role of Semantic Technologies in a New Service

Elservier

■概要

- 研究データセットのショーケースサービス構築についての紹介

■課題

- 複数リポジトリに登録されているデータセットに機関名が紐付いていない．また，自由記述のため表記揺れがある

■解決手段

- 学術論文とデータセットをつなぐThe Scholix frameworkにElsevierのシステムを接続し，入手したデータを利用してKGを構築
- KGを利用して機関名の曖昧性を解消

■実績

- Mendeleyのサイトでデータセット検索を提供中
  (ショーケースの提供有無は不明)

How to Maintain a Linked Data Cloud in a Deployed Semantic Portal

Aalto University, University of Helsinki

■概要

- WW2データをLODで公開した経験から, データ保守に関する知見の紹介

■課題

- 元データの追加時に生じる, データ変換, エンティティリンキング, グラフへの変更の反映など

■解決手段

- LODを直接編集するのではなく，ドメイン専門家が元データセットに対して作業を行い，再度LOD化する
- 長期的な保守は，ドメイン知識を持った管理組織が行う．そのためには，IT専門家でない人が簡単に使える管理アプリが必要

■実績

- WarSampoを公開中
  https://www.sotasampo.fi/en/

Linking Media: adopting Semantic Technologies for multimodal media connection

Vilynx, カタルーニャ工科大学

■概要

- KGを利用したメディアコンテンツの検索・推薦サービスの構築について紹介

■課題

- メディアコンテンツに対するラベル付コスト

■解決手段

■実績

- Vilynxプラットフォームでサービス中． 2万件/日のコンテンツをインデクシング

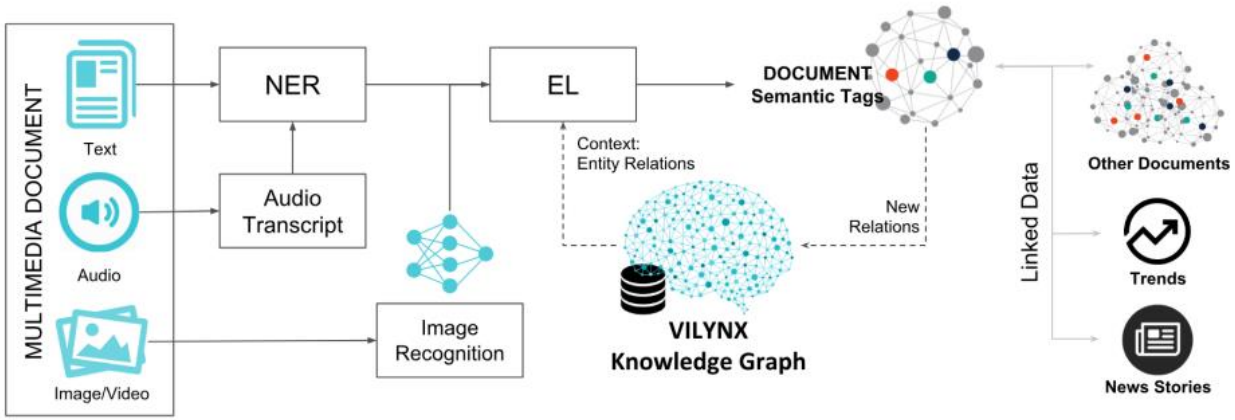


Fig. 1. Scheme of Vilynx's multimodal document tagging framework.

Product Search using Linked Data

Semaku, Nexperia

■概要

- LDを利用した半導体製品の検索サービスの構築ついて紹介
- Nexperiaは1万種以上の製品を製造販売している

■課題

- ニーズに対応してSPARQLクエリを更新し，徐々に複雑化．パフォーマンスやメンテナンス性の低下が顕著になってきた

■解決手段

- 単発クエリを複数のCONSTRUCTクエリに分割し，結果をマージする方式に変更
- LODを直接編集するのではなく，ドメイン専門家が元データセットに対して作業を行い，再度LOD化する
- JSON-LD Framingを利用してグラフを階層構造化し，ローカルのURIを付与してインデックスに登録

■実績

- Nexperia Data Portalでサービス中
  http://www.data.nexperia.com/

1.3.1 Embeddings and Deep Learning

Fine-grained Evaluation of Rule- and Embedding-based Systems for Knowledge Graph Completion

University of Mannheim

■概要

- KGCタスクにおける”きめ細かな”モデル評価手法に関する提案

■ 先行研究との差分

- KGCタスクにおけるルールベース型アプローチと埋込み型アプローチを比較評価するベースラインを作成した

■ 技術ポイント

- データセットの特徴に基づいて分割したものを利用することで，従来より”きめ細かな”粒度でモデル比較が可能になった

■ 評価

- 規則性*1に着目してデータセットを分割，両アプローチの長所・短所を比較
  *1: Symmetry, Inverse, Equivalence, Subsumption
- アンサンブルモデルで”良いとこ取り”

■ 議論

- 提案評価手法の一般化

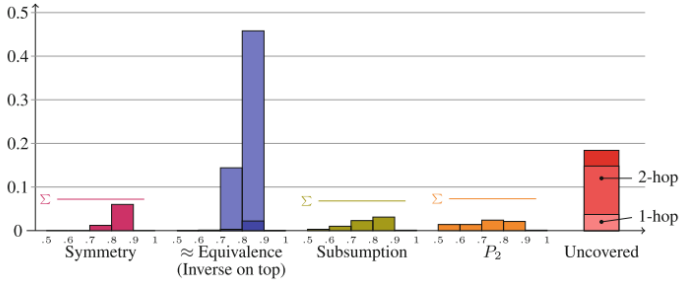


図2: FB15kのルールカバレッジ

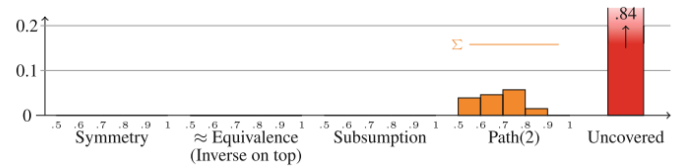


図3: FB15k-237のルールカバレッジ

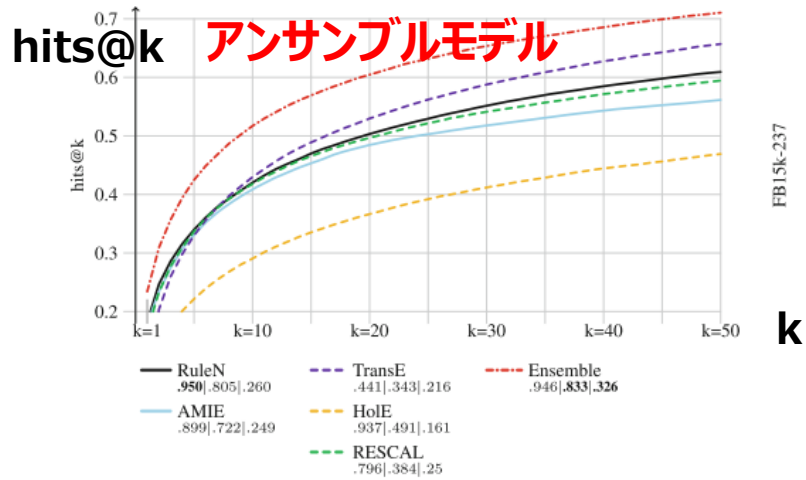


図4: FB15k-237に対する各モデル，及びアンサンブルモデルの比較評価

Aligning Knowledge Base and Document Embedding Models using Regularized Multi-Task Learning

University of Zurich, Zhejiang University,
Alibaba-Zhejiang University Joint Institute of Frontier Technologies

■概要

- 知識グラフ埋込みモデルと文書埋込みモデルのアラインメント手法を提案

■ 先行研究との差分

- 先行研究なし

■ 技術ポイント

- 知識グラフのエンティティと文書の対応関係を維持するように，知識グラフ埋込み，文書埋込みを各々学習する（マルチタスク学習）
- アラインメントの損失関数が不要

■ 評価

- 知識グラフ埋込みモデル，文書埋込みモデルを個別に学習した場合と比較して精度の遜色なし

■ 議論

- エンティティと文書を新規追加した場合の埋込みモデル更新は今後の課題

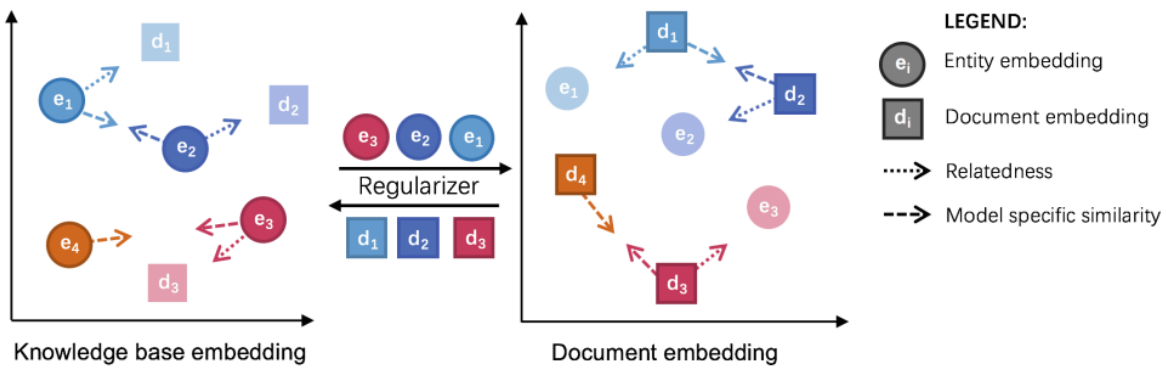


Fig. 1. Illustration of the intuition behind KADE. (Color figure online)

Inducing Implicit Relations from Text using Distantly Supervised Deep Nets

IBM Research AI, Universidad Politcnica de Madrid, University of Bari

■ 概要

- コーパス中に明示されていない暗黙の関係性を抽出する

■ 先行研究との差分

- コーパスへの教示を行うことなく関係抽出を行う

■ 技術ポイント

- Distant supervisionを利用し，既知の関係知識からコーパスにアノテーションし学習データを生成する
- そのままだとノイズデータが増えるため，まずunary関係を抽出してKBを構築し，それに対してどのエンティティを割り当てると正しいbinary関係になるかを学習する

■ 評価

- ISWC SWC2017優勝ﾁｰﾑを超える精度
- Precision-RecallカーブでRecallが低下しない(unary/binary単独モデルと比較)

■ 議論

- ―

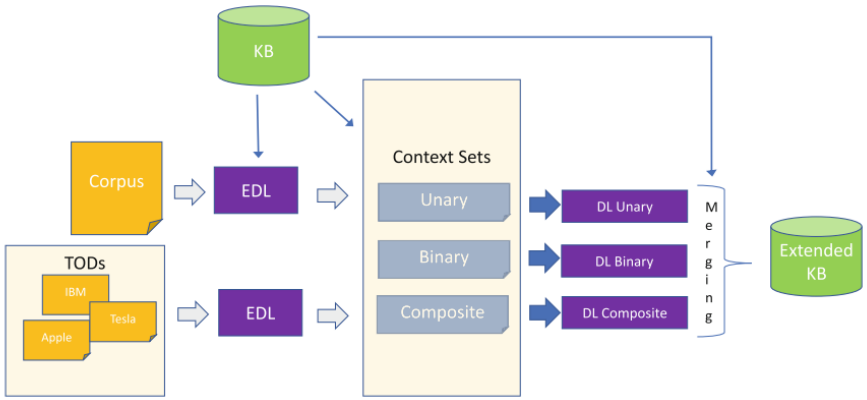


Fig. 1. Socrates architecture

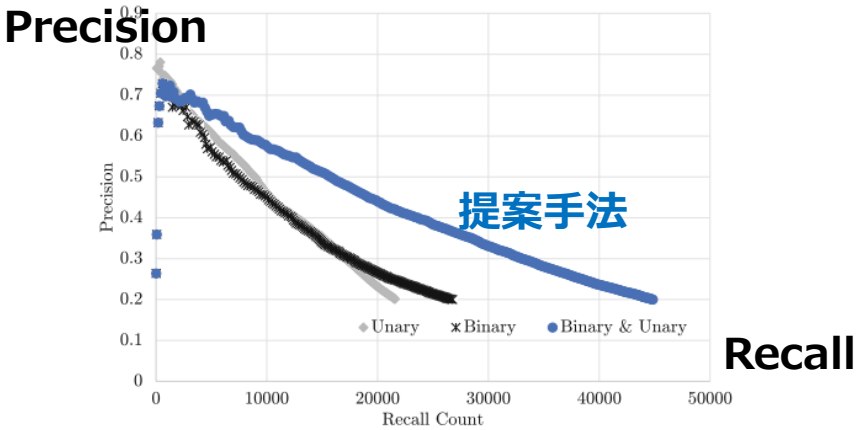


Fig. 5. Precision recall curves for KBP on CC-DBP

お疲れ様でした！