

学生・若手研究者のためのBERTワークショップ（2回目）
～BERTを使った自然言語処理の開発を
2回のワークショップで体験！～

理化学研究所
革新知能統合研究センター
言語情報アクセス技術チーム

オープニング



本日のスケジュール

- 質疑応答（40分）
 - Slack上の質問への解説
 - 新規質問があればslackへ。答えられるものは答えます。
- 体験報告(30分)
 - 結果提出された人の中から 5 名の方に体験報告（各自5分程度）
 - 時間があればその他の方からの体験報告
- 森羅プロジェクトの説明（20分）
 - 知識構造化プロジェクトの紹介
 - メンバー募集



次回のワークショップ



- 属性抽出のタスクをBERTで解く
 - 9月中旬～下旬
- 他の希望があれば、slackに意見を

学生・若手研究者のためのBERTワークショップ（2回目）
～BERTを使った自然言語処理の開発を
2回のワークショップで体験！～

理化学研究所
革新知能統合研究センター
言語情報アクセス技術チーム

体験報告



体験報告

- 5 名の方の体験報告（リーダーボード提出者）
 - みうら
 - Yuanzhi Ke
 - Akira Ogawa
 - Team久山田
 - NLPオールドタイマーズ
- 時間があれば、その他の方からの体験報告

森羅

SHINRA

理化学研究所 革新知能統合研究センター

「拡張固有表現＋Wikipedia」構造化データ

BERT Workshop 2nd day: 2022/8/10

<http://shinra-project.info/>

「森羅」ホームページ

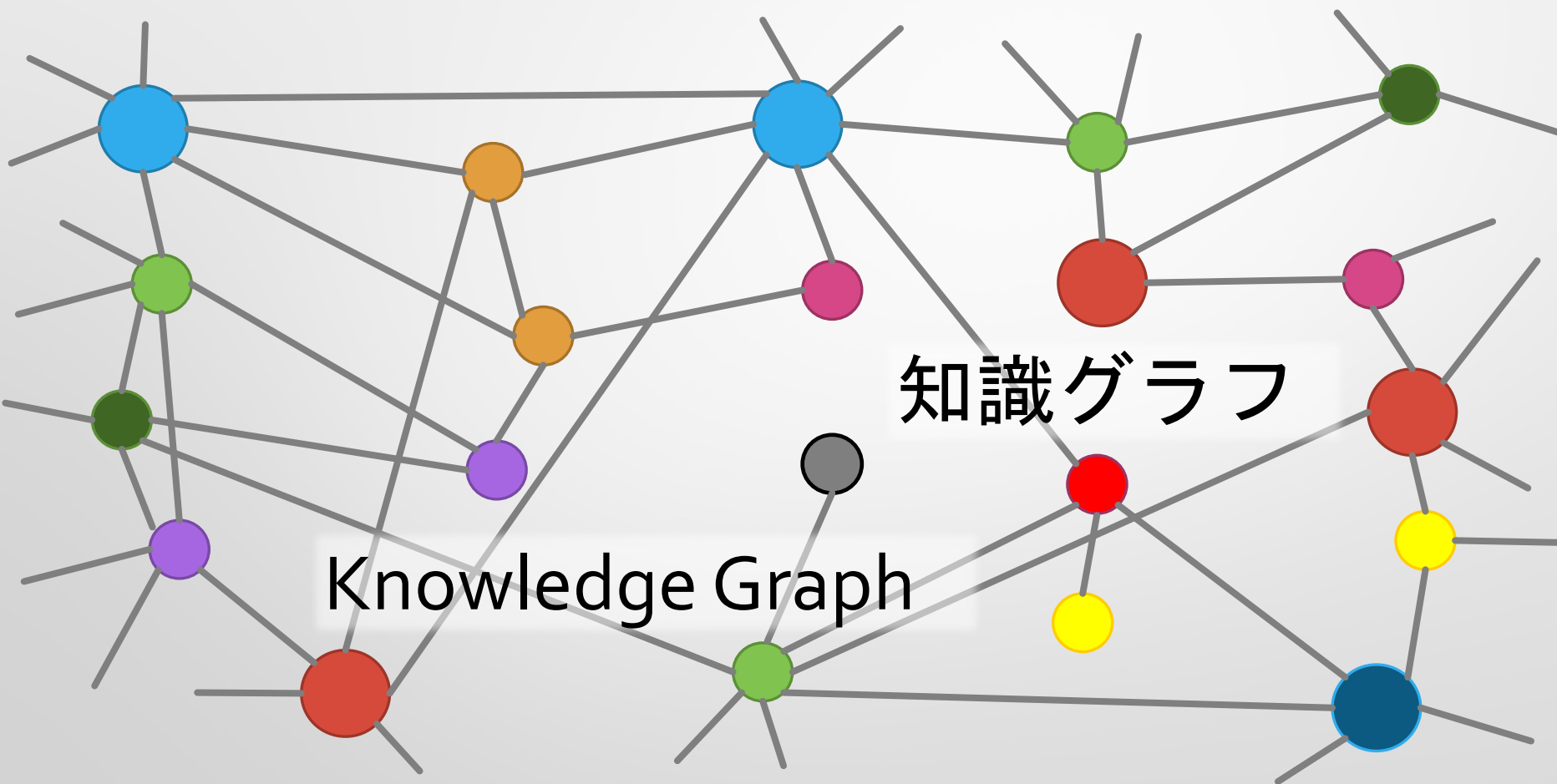




森羅

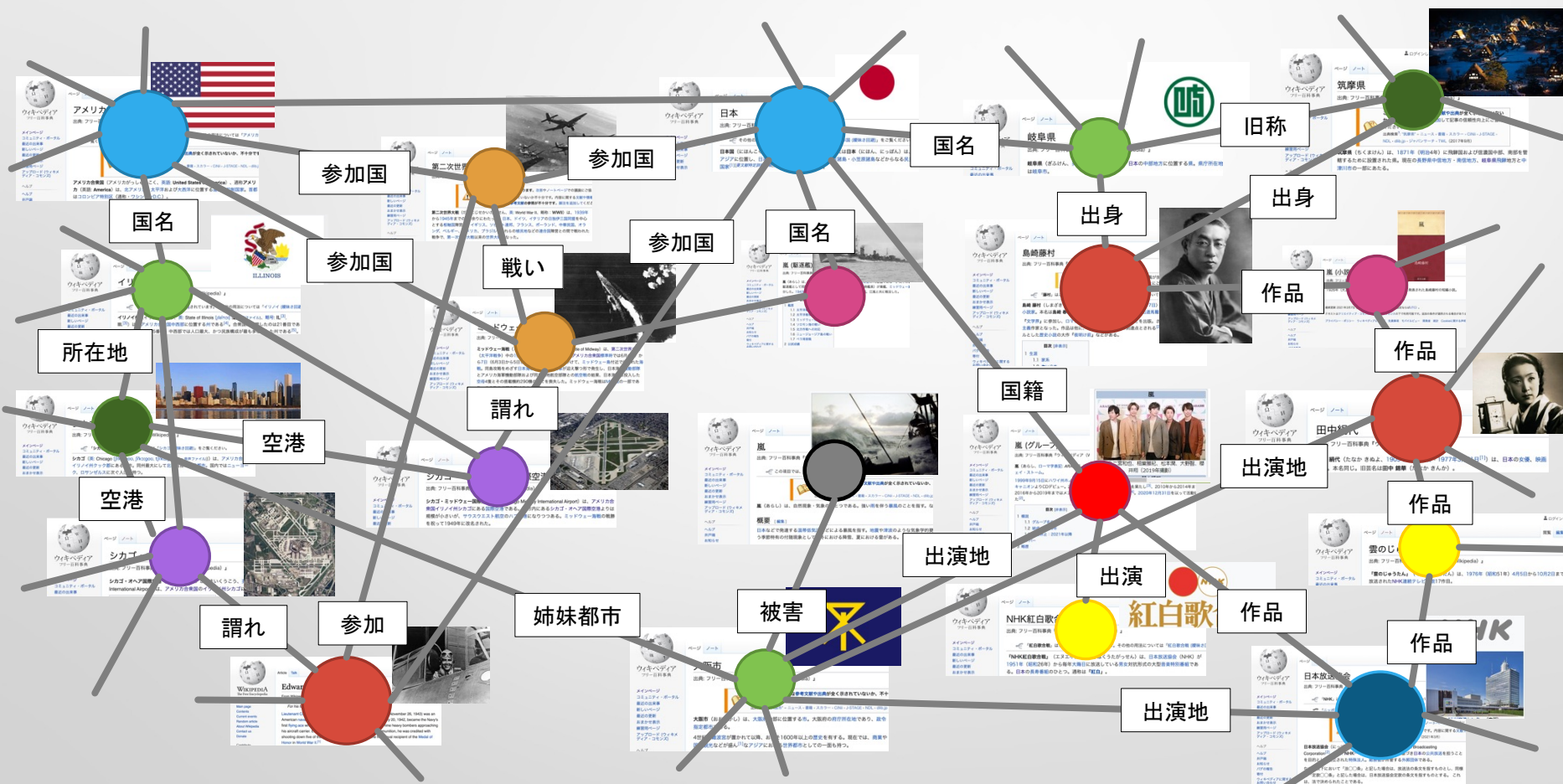
=

Knowledge Graph
知識グラフ



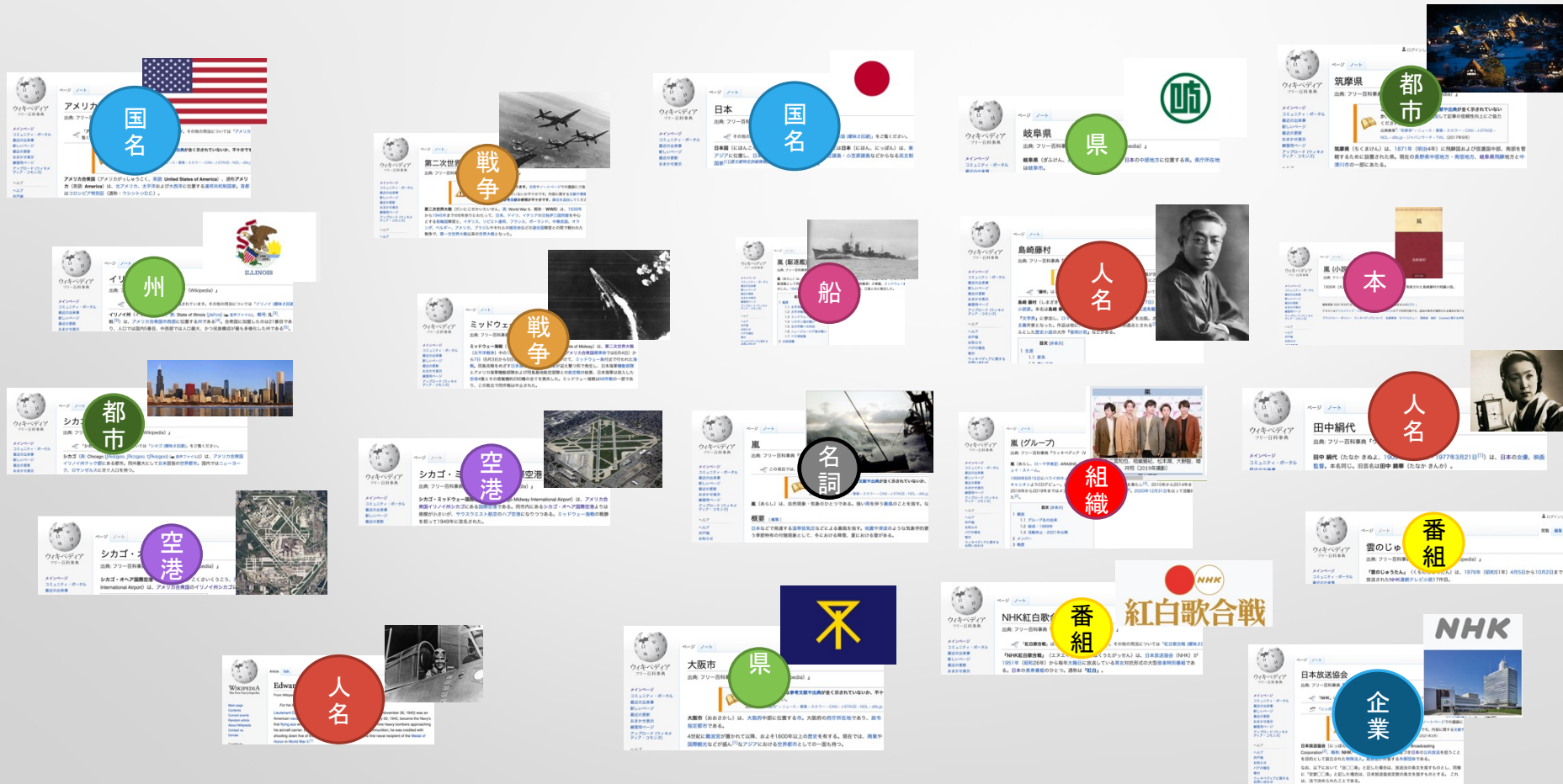


Knowledge Graph 知識グラフ





ステップ1 (分類)





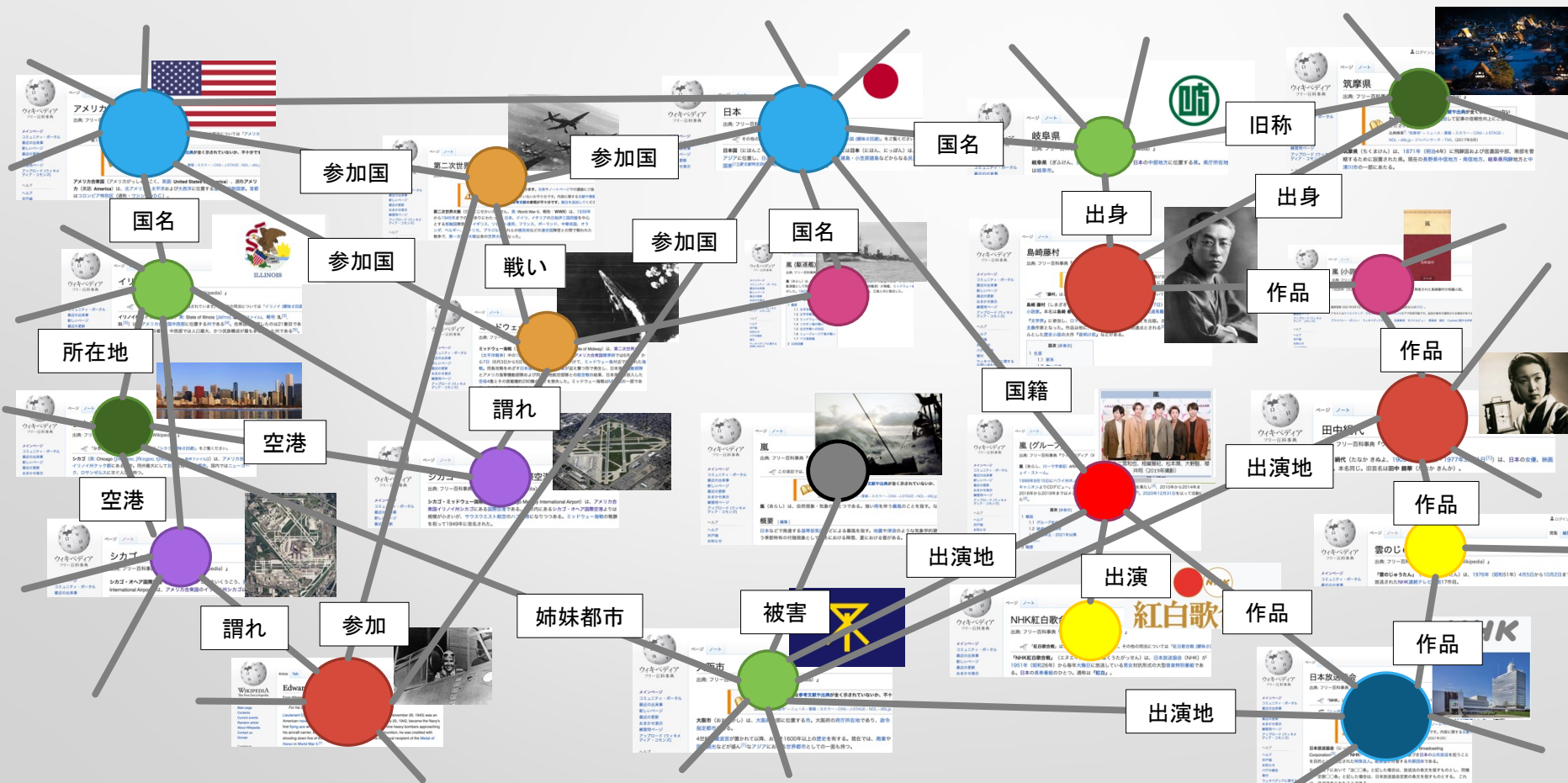
ステップ2 (属性値抽出)

属性	属性値
国	アメリカ合衆国
所在地	イリノイ州
設立	1837年3月4日
人口	2695万人
人口データの年	2010年
空港	オヘア空港、ミッドウェー空港
...	...

属性	属性値
本名	島崎春樹
生年月日	1872年3月25日
国籍	日本
出身地	筑摩郡（現在の岐阜県）
地位職業名	小説家
作品	若菜集、嵐、夜明け前。。
...	...



ステップ3 (リンクング)



知識グラフを作ろう！

3つのステップ

ステップ1（分類）

各Wikipediaページを約220種類の拡張固有表現に分類
（「島崎藤村」は人名！）

ステップ2（属性値抽出）

固有表現定義にある属性値をページから抽出
（「島崎藤村」の「作品」には「嵐」がある！）

ステップ3（リンクング）

抽出した属性値を該当するWikipediaページに紐付け
（「嵐」はWikipediaページの「嵐（作品）」のこと！）



RbCC

Resource by Collaborative Contribution
共働による知識構築





協働による知識構築

Resource by Collaborative Contribution (RbCC)



- 評価型ワークショップを実施
- 単に性能を競い合うだけではない
- 参加システムがリソース作成に直接貢献
 - 例えば、10チーム中8チームが正しいと
いったものは正しいとする (Ensemble
Learning)
 - 適切な人手チェックを入れデータを拡張
(Active Learning)
 - 拡張した教師データで再度タスクを実行
(Bootstrapping)
- すべての出力データは参加者で共有する
- 統合されたデータは一般に公開する





森羅2018 結果(Micro F)



System	Person	Company	City	Airport	Compound
TUT	20	41	28	72	
OCU	19				
NUT					42
Sansan		30			
Fuji Xerox	31		43	42	39
Toppan		33		35	
Unisys	44	53	42	67	47
AIP	36	38	46	71	46
Ensemble	48	61	58	87	65
UP	+4	+8	+12	+15	+18



目的





森羅プロジェクトの最終目標



信頼される人工知能

単に答えを示すだけではなく、
答えの根拠を人が理解できる
説明の形で提示する。

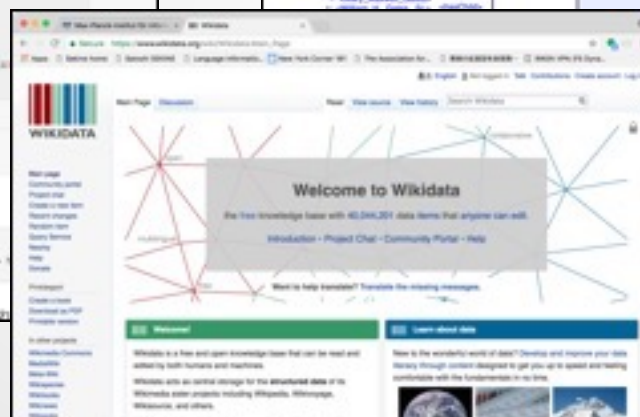
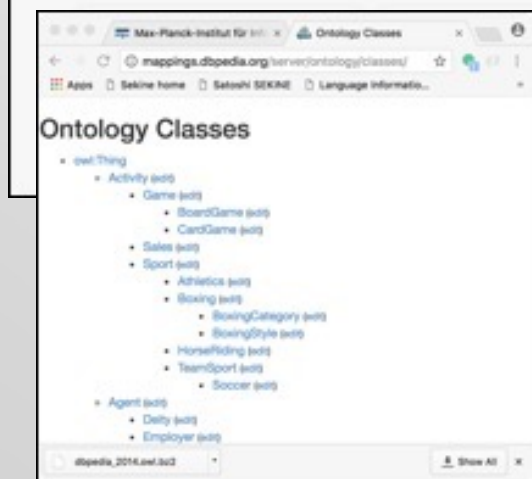
人工知能の普及に新しい展開



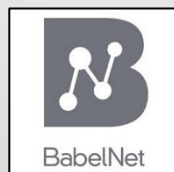
構造化された世界知識
「森羅」



既存の知識グラフ



とにかく、
汚い



現在ある知識ベースは
自然言語処理利用に耐えられない



知識グラフを作ろう！



Name	Facility	Product	Natural Object
Person Individual Animal Individual Animal Other Organization International Organization State Organization Family Ethnic Group Nationality Ethnic Group Other Sports Organization Sports Federation / Sports Team Sports League Sports Organization Other Academic Person Nonprofit Organization / Company Company Group Judicial Person Other Political Organization Government / Political Party Cabinet Military Political Organization Other Organization Other Location GPS City / Province / Country / GPS Other Continent Region Country Region Region Other Geographical Region Sea / Mountain / Island / River Lake Sea Bay Geographical Region Other Astronomical Object Star Planet / Constellation Astronomical Object Other Address Postal Address Address Other Location Other	Facility Part Room Archaeological Place Tomb Archaeological Place Other FOL Village / Hamlet / Castle / Estate Public Institution / Accommodation Judicial Institution / School Research Institute / Market / Power Plant Park / Shopping Complex Sports Facility Museum / Zoo / Amusement Park Theater / Wedding Place FOL Other Transport Facility Car Ship Station Airport / Port / Transport Facility Other Law Railroad Road Canal Water Route / Tunnel Bridge Law Other Facility Other Event Election Election Religious Festival / Competition / Conference Occasion Other Incident War Incident Other Language Natural Phenomenon Language Other Earthquake Natural Phenomenon Other Event Other Color Natural Color Color Other Disease Animal Disease Disease Other	Value Work Musical Instrument / Clothing / Money / Jewelry / Drug / Weapon / Book / Award / Dissertation / Officer / Service / Class / Character / ID Number Game Digital Game / Game / Game Other Software Vehicle Car / Train / Aircraft / Spaceship / Ship / Vehicle Other Food Meat Food Other Drink Food Other Furniture / Bedstead / Program / Mirror / Store / Music / Book / Art Other Printing Newspaper / Magazine / Printing Other Doctrine Method Culture Religion Academic / Theory Plan Doctrine Method Other Rule Treaty Law / Rule Other Title Position / Vocation / Title Other Rank Language Natural Language / Language Other Unit Currency / Unit Other Product Other Virtual Address Channel / Phone Number / Email / URL Virtual Address Other Name Other	Element Compound Mineral Living Thing Fungus / Mollusc / Arthropod / Insect / Fish / Amphibia / Reptile / Bird / Mammal / Plant / Living Thing Other Living Thing Part Animal Part / Plant Part / Living Thing Part Other Natural Object Other Numex Money / Stock / Index / Point / Percent / Multiplication / Frequency / Age / School Age / Ordinal Number / Rank / Latitude / Longitude / Measurement Physical Event / Space Volume / Weight / Speed / Density / Temperature / Color / Season / January / Season / Magazine / Measurement Other Count N Person / N Organization / N Location / N Facility / N Product / N Event / N Natural Object / N Animal / N Plant / N Natural Object Other / N Name Other Timex Event Time / Date / Day / Week / Era / Times Other Period Period / Time / Period / Day / Period / Week / Period / Month / Period / Year / Period Other CONCEPT IGNORED

Wikipediaの記事を拡張固有表現に分類し
定義された属性値を抽出して構造化、リンク作成

Resource by Collaborative Contribution

Since 2017

評価WSを通して協働で知識作成

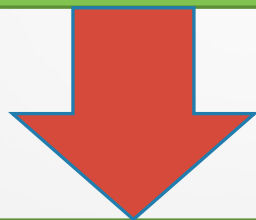


精度向上と共に...

残っている大きな問題！



Wikipediaは更新され続けていく



森羅を更新し続けるために
仕組みを(半)自動化することが必要



(半)自動的なアップデート



- 過去の「森羅データ」を教師として利用
 - 森羅2019を教師としてW2021を(半)自動で構造化
 - 森羅2021を教師としてW2023を(半)自動で構造化
 - 森羅2023を教師としてW2025を(半)自動で構造化
 - ...
- 3つのステップを一気に実施
 - 分類、属性抽出、リンキングの複合タスク
 - 相乗効果／End-to-Endで精度向上の可能性

知識グラフがあると

物知り博士



第二次世界大戦に由来した名前を持つ2つの空港がある米国の都市はどこ？

第二次世界大戦の英雄に由来するオヘア空港と戦場名に由来するミッドウェー空港があるシカゴ！

雑談対話

おじいちゃんの好きな田中絹代が出演した1957年の「嵐」って映画の原作は、島崎藤村の小説なんだって



情報アクセス

教育支援

営業支援

認知症予防

介護福祉

多言語情報アクセス

他にも。。。

特定応用展開

観光

外国語教育

ビジネス

特許検索・分析

法律文書解析

オンライン医療

健康自己管理



森羅2022タスクについて





タスクスケジュール



- キックオフミーティング／データ公開：
2022年5月12日10:30-12:30
- 実行結果の提出：2022年10月頃
- 評価結果の返却：2022年11月頃
- ワークショップ：2022年12月



評価



- リーダーボード
 - 3タスクのリーダーボード
 - Public評価（提出された内の一部で評価。10月頃まで継続）
 - Private評価（Publicとは違うデータで評価、10月以降に実施）
- 本評価
 - Wikipedia2021の全データに対しての結果に対して評価を行う
 - 実行には大きな計算機リソースと時間が掛かる
 - 実施困難な場合などには相談に乗ります



HP、コミュニティ



- 森羅プロジェクトHP
<http://shinra-project.info>
- 森羅2022ホームページ
<http://shinra-project.info/shinra2022/>
- slack (<https://shinra2022.slack.com/>)
参加はこちら https://join.slack.com/t/shinra2022/shared_invite/zt-1dlrhtpty-eCBovUEWpJj1k4Uhugh4Fw
- ML
 - shinra2022-all _at_ [googlegroups.com](https://groups.google.com/g/shinra2022-all) (タスク関係のアナウンス)
参加はこちら <https://groups.google.com/g/shinra2022-all>
 - shinra2022-info _at_ [googlegroups.com](https://groups.google.com/g/shinra2022-info) (オーガナイザーへの質問)



メンバー募集



- 学生バイト／インターン

- 目安とする対象：B4-D1
- 週2日程度のアルバイト
- 期間を限定してインターン
- 時給などは理化学研究所の規定による
- RAIDEN(超大型GPUマシン) 利用可能

- 社会人

- 自由時間に協力
- 会社からの兼業許可

共に
slackで“Satoshi Sekine”宛に
ダイレクトメッセージ
をお送りください

面談をさせていただきます

分類システム作成の希望者は
システム提出をお願いします

他の形での参加を希望する方は
メッセージにその旨をご記入下さい



次回のワークショップ



- 属性抽出のタスクをBERTで解く
 - 9月中旬～下旬
- 他の希望があれば、slackに意見を