ミーティングは13:00に開始します

http://shinra-project.info/

本日のスケジュール

13:00	オープニング、森羅全体説明	関根
13:15	森羅2023タスクの説明	安藤
13:30	森羅公開データの説明	三浦
13:45	リーダーボードの説明	門脇
13:45	デモアプリ製作トラックの説明	宇佐美
14:00	質疑応答	
14:10	*休憩*	
14:15	招待講演	
	「Bing対話型検索とGPTモデル」 （元 Microsoft Bing対話型検索プロダクトマネージャー）	鈴木久美先生（online）
14:55	クロージング	
15:00	閉会	

目的

森羅プロジェクトの最終目標

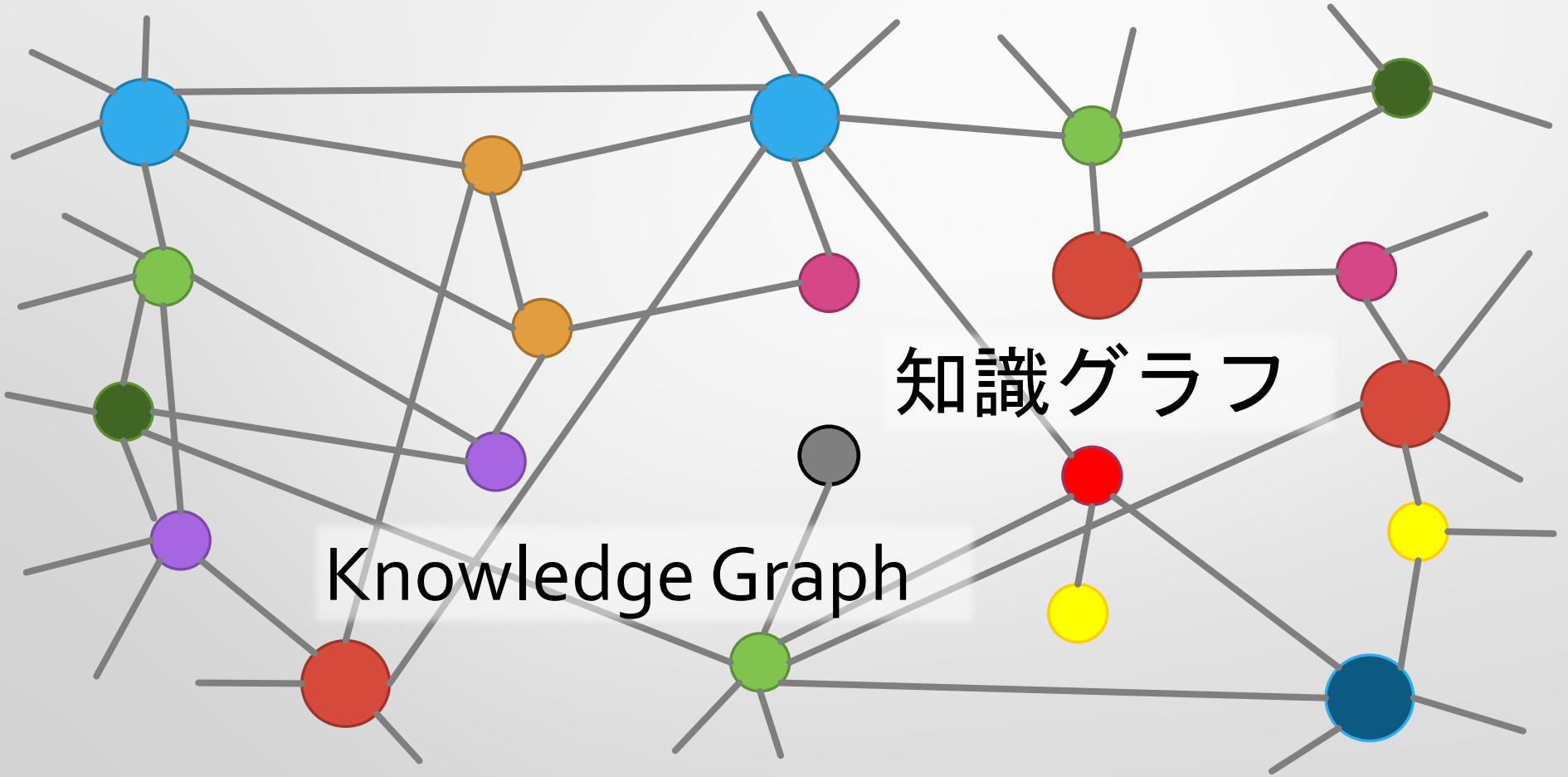
信頼される人工知能

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」

森羅 ＝ Knowledge Graph 知識グラフ



森羅：Wikipedia構造化プロジェクト

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！

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

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（リンキング）

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！）

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！）

SHINRA2020-ML
SHINRA2021-ML
SHINRA2022
80~93

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！）

SHINRA2018
SHINRA2019
SHINRA2020-JP
SHINRA2022
50~90

ステップ３（リンキング）

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！）

SHINRA2021-LinkJP
SHINRA2022
80~90

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

３つのタスクを同時に行う理由

相乗効果の期待

- End-to-Endのシステムで精度向上を目指す
- 前段のタスクがNベストを出力し再順位付けする
  - 属性値抽出タスクの結果
    →取れた属性から 分類の間違えを修正
  - リンクタスクの結果
    →リンク先のカテゴリーを用いて属性を修正

単独タスク参加のためにBaselineシステムを公開

森羅：Wikipedia構造化プロジェクト

RbCC

Resource by Collaborative Contribution

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ー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**

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

他にも。。。

多言語情報アクセス

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：Wikipedia構造化プロジェクト

森羅2023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：
  2023年5月18日13:00-15:00
- 中間報告会：2023年7月頃
- 最終報告会：2023年12月

評価

- リーダーボード
  - 「３タスクのそれぞれ」と「End-to-End」の４つのリーダーボード
  - 評価（リーダーボード対象データで評価）
    - 2022よりデータの精度を向上させたデータで、新規のリーダーボードを用意

- 本評価
  - Wikipedia2021の全データに対して、３タスク／End-to-Endを実施
  - それぞれの結果を提出していただき、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ため、参加者に準備が必要
  - 参加希望の方はあらかじめ相談下さい

HP、コミュニティー

- 森羅プロジェクトHP

  http://shinra-project.info

- 森羅2023ホームページ

  https://2023.shinra-project.info//

- slack（https://shinra2022.slack.com/）