学生・若手研究者のためのBERTワークショップ２
（固有表現タスク　２回目）

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

本日の概要

- 14:00-14:05　オープニング
- 14:05-14:35　属性値抽出タスクの紹介（三浦）
  - 森羅の１タスクである「属性抽出タスク」を固有表現抽出タスクとの関連から紹介する
- 14:35-15:35　体験報告&第1回交流会（司会：渋木）
  - 固有表現システム開発の体験報告（３チーム）
  - 交流会（関連研究の紹介：2名）
- 15:35-16:00　森羅プロジェクト紹介＆クロージング

森羅Slack ：shinra2022.slack.com
チャンネル：#bert_workshop2022
招待リンク：http://shinra-project.info/shinra2022/shinra2022_slack_invite

学生アルバイト募集

このWSのその先に

- 森羅プロジェクトでアルバイトを募集しています
  - 対象：Ｂ４~Ｄ１を対象
  - 形態：アルバイト（１年程度、週に１５時間前後、リモート可）
    インターン（２～３ヶ月程度集中的に。リモート相談）
  - 業務内容：森羅プロジェクトのシステム開発、データ開発
    指導教官との相談の上、卒論、修論とすることも可能
    理研のRIDEN（超巨大GPUマシン）を利用可能

- 社会人の兼業希望者も応相談

相談、希望は下記までメイルを
satoshi.sekine@riken.jp


クロージング

本日の概要

- 14:00-14:05　オープニング
- 14:05-14:35　属性値抽出タスクの紹介（三浦）
  - 森羅の１タスクである「属性抽出タスク」を固有表現抽出タスクとの関連から紹介する
- 14:35-15:35　体験報告&第1回交流会（司会：渋木）
  - 固有表現システム開発の体験報告（３チーム）
  - 交流会（関連研究の紹介：2名）
- 15:35-16:00　森羅プロジェクト紹介＆クロージング

森羅Slack ：shinra2022.slack.com
チャンネル：#bert_workshop2022
招待リンク： https://join.slack.com/t/shinra2022/shared_invite/zt-14qkpf21i-lQNKlToaIOU5We7xlZBqfQ

森羅プロジェクト

- Wikipediaの構造化プロジェクト
  - 「説明できる人工知能」を作るために知識グラフを構築
  - 協働による知識構築
    - リーダーボードを作り、参加者とともに実施

- 3つのタスク
  - Wikipediaページを約200種類のカテゴリーに分類
  - ページの属性情報を抽出
  - 抽出した情報をリンキング

- 進捗
  - 2018年からタスクを実施。今年度、来年度は3つのタスク

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！）

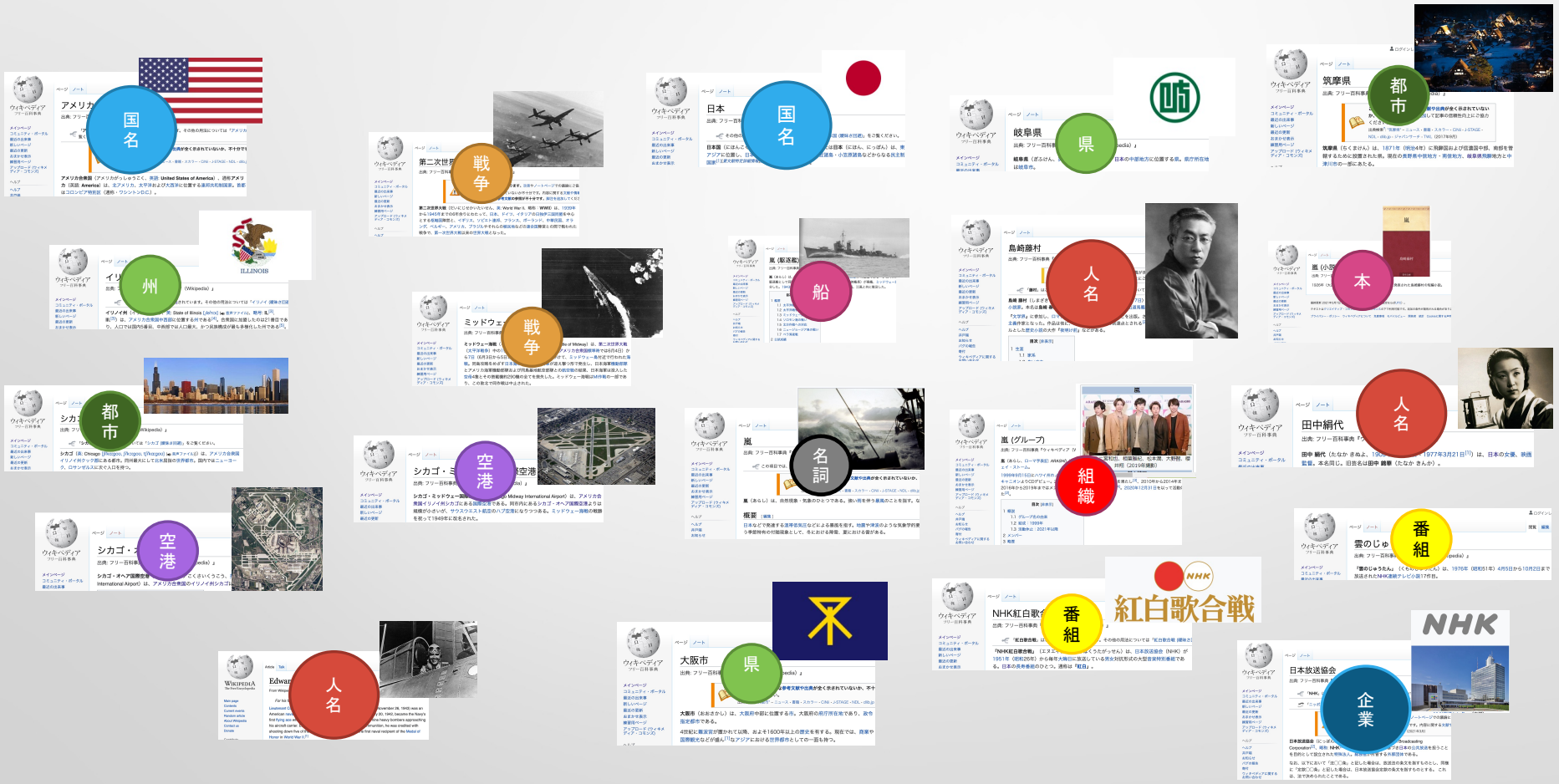
ステップ３（リンクの紐付け）

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！）

Wikipediaページ

（日本語は100万、英語は600万）

ステップ１（分類）

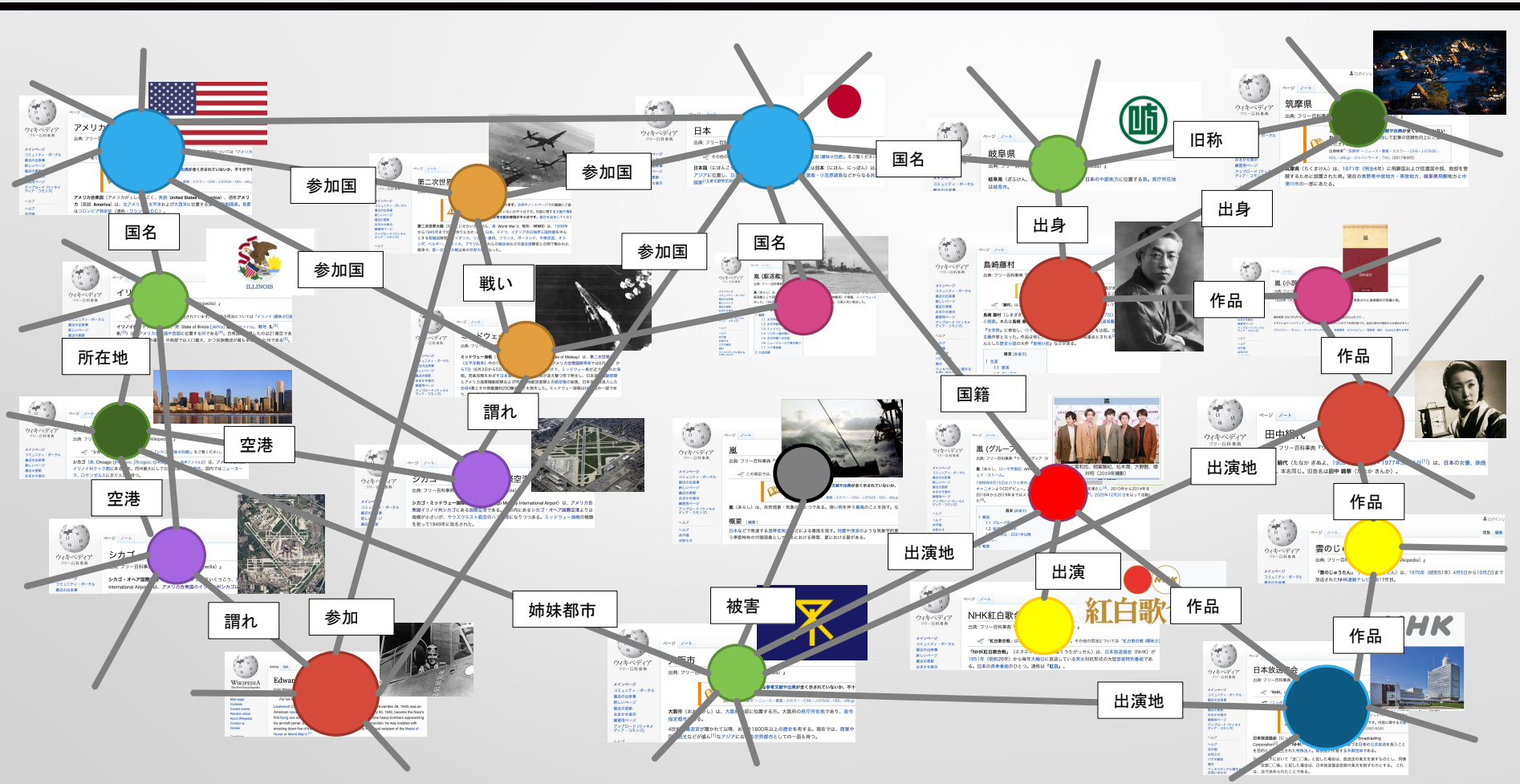


ステップ２
（属性値抽出）

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	オヘラ空港、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
（リンキング）


国名
参加国
参加国
国名
参加国
戦い
国名
所在地
空港
空港
謂れ
謂れ
参加
姉妹都市
被害
出身
旧称
出身
作品
作品
国籍
出演地
出演
出演地
作品
作品
作品
出演地

既存の知識グラフ

Freebase™

BabelNet

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

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**

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

今後のタスク

- リーダーボード
  - 分類、属性値抽出、リンキングとも締切なく継続
- 分類タスク（２０２２年本評価）
  - 日本語Wikipedia全体に対してシステムを走らせ、結果を提出
  - １１月１４日締切
- 他の２０２２年本評価
  - 属性値抽出、リンキングの本評価は中止
- ２０２３年の本評価
  - ３つのすべてのタスクの本評価を行う

HP、コミュニティー

- 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

  http://shinra-project.info/shinra2022/shinra2022_slack_invite

学生アルバイト募集

このWSのその先に

- 森羅プロジェクトでアルバイトを募集しています

  対象：Ｂ４~Ｄ１を対象

  形態：アルバイト（１年程度、週に１５時間前後、リモート可）
  　　インターン（２～３ヶ月程度集中的に。リモート相談）

  業務内容：森羅プロジェクトのシステム開発、データ開発
  　　指導教官との相談の上、卒論、修論とすることも可能
  　　理研のRIDEN（超巨大GPUマシン）を利用可能

  - 社会人の兼業希望者も応相談

相談、希望は下記までメイル
satoshi.sekine@riken.jp
または、slackで個人メッセージを