学生・若手研究者のためのBERTワークショップ（１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

本日の概要

- 今、流行りのBERTって何？（30分）
  - レクチャーによる説明
- BERTをとりあえず動かしてみよう(60分）
  - Google Colabを使って自分で動かしてみる
    - Googleアカウント、Google Drive上に600MB必要
  - コードの説明
- リーダーボード、コミュニティーの紹介（15分）
- 森羅プロジェクトの説明、今後の説明（5分）

事前知識がない方は
難しいと思っても大丈夫！
ビデオで再学習できます

Slackで連絡を
行います。
只今、zoomの
チャットで招待
リンクを共有

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-1dlrhtpty-eCBovUEWpJj1k4Uhugh4Fw

チームメンバー募集

このWSのその先に

- （BERTを使って）森羅の研究に取り組みたい学生を募集
  - 最大3名（だいたいB4~D1を対象）
  - チームにアルバイト、インターンとして参加
  - 期限はまず今年度末（相談の上、来年度の延長も可能）
  - 2回目（8月10日）に応募方法をアナウンス

本日の講師の中山功太さんも、2017年のインターン学生

  - 社会人の希望者があれば、応相談

リーダーボード作成等の門脇さんも、社会人参加者

本日の講師：中山功太

学歴

2014年4月 – 2018年3月(学部), 2018年4月 - 2020年3月 (修士課程)
豊橋技術科学大学 システムAIラボ(石田研)
2020年4月 – 現在(博士課程)
筑波大学 ヒューマンコンピューテーション研究室(馬場研)

職歴

2018年8月 – 2020年10月
理研 言語情報アクセス技術チーム(関根チーム) 研究パートタイマー
2020年11月 – 現在
同上 リサーチアソシエイト

論文(査読付き)

- Kouta Nakayama, Shuhei Kurita, Akio Kobayashi, Yukino Baba, and Satoshi Sekine. 2021. Co-Teaching Student-Model through Submission Results of Shared Task. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4525–4535, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Satoshi Sekine, Kouta Nakayama, Maya Ando, Yu Usami, Masako Nomoto and Koji Matsuda, SHINRA2020-ML: Categorizing 30-language Wikipedia into fine-grained NE based on "Resource by Collaborative Contribution" scheme, In Proceedings of the 3rd conference on the Automated Knowledge Base Construction (AKBC 2021), 2021.
- Satoshi Sekine, Akio Kobayashi, Kouta Nakayama, SHINRA: Structuring Wikipedia by Collaborative Contribution, In Proceedings of the 1st conference on the Automated Knowledge Base Construction (AKBC 2019), 2019.

学生・若手研究者のためのBERTワークショップ（１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

クロージング

本日の概要

- 今、流行りのBERTって何？（30分）
  - レクチャーによる説明
- BERTをとりあえず動かしてみよう(60分）
  - Google Colabを使って自分で動かしてみる
    - Googleアカウント、Google Drive上に600MB必要
  - コードの説明
- リーダーボード、コミュニティーの紹介（15分）
- 森羅プロジェクトの説明、今後の説明（5分）

森羅プロジェクト

今回のWSのサンプルデータ

- 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

今回のタスク

- 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）

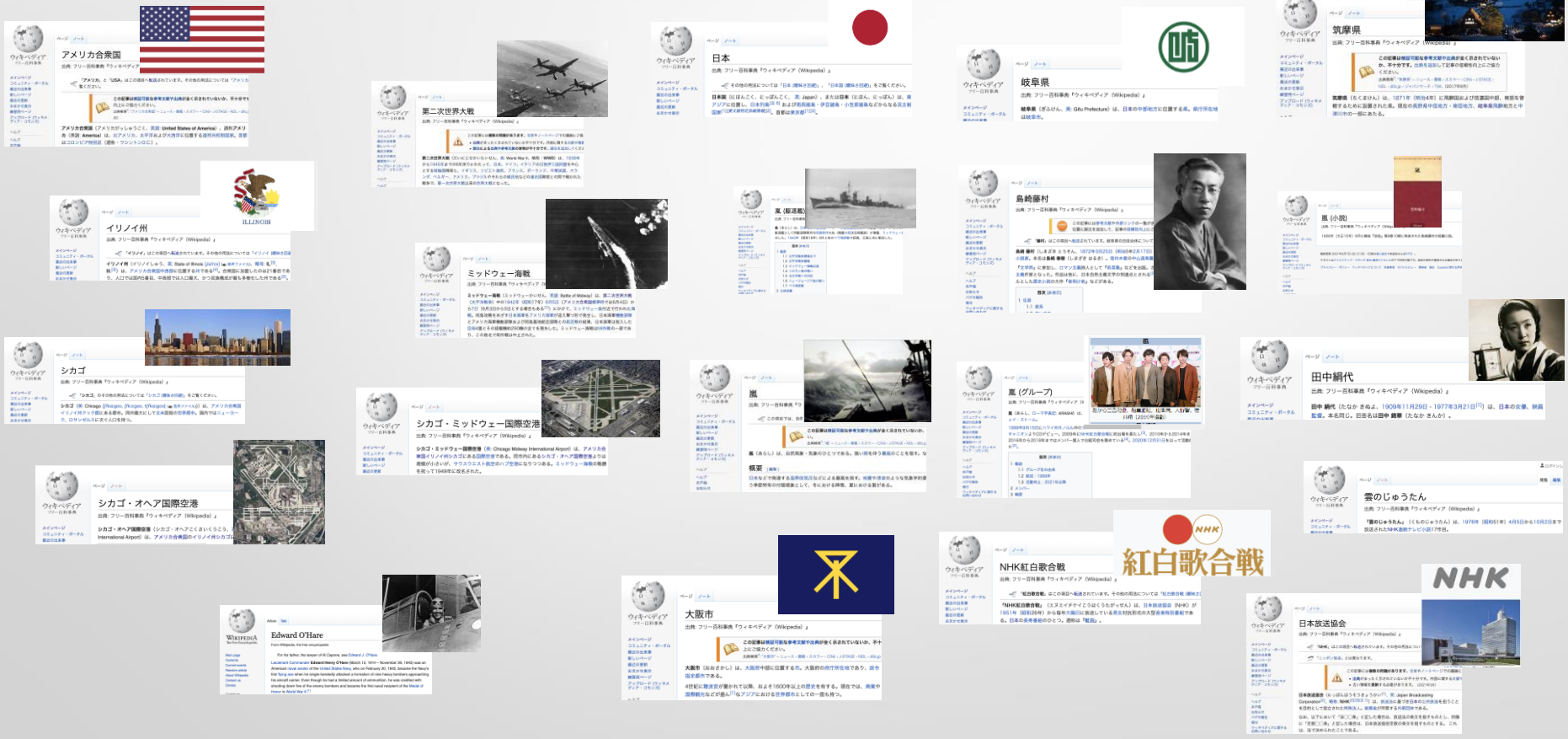
固有表現定義にある属性値をページから抽出
（「島崎藤村」の「作品」には「嵐」がある！）

ステップ３（リンクの紐付け）

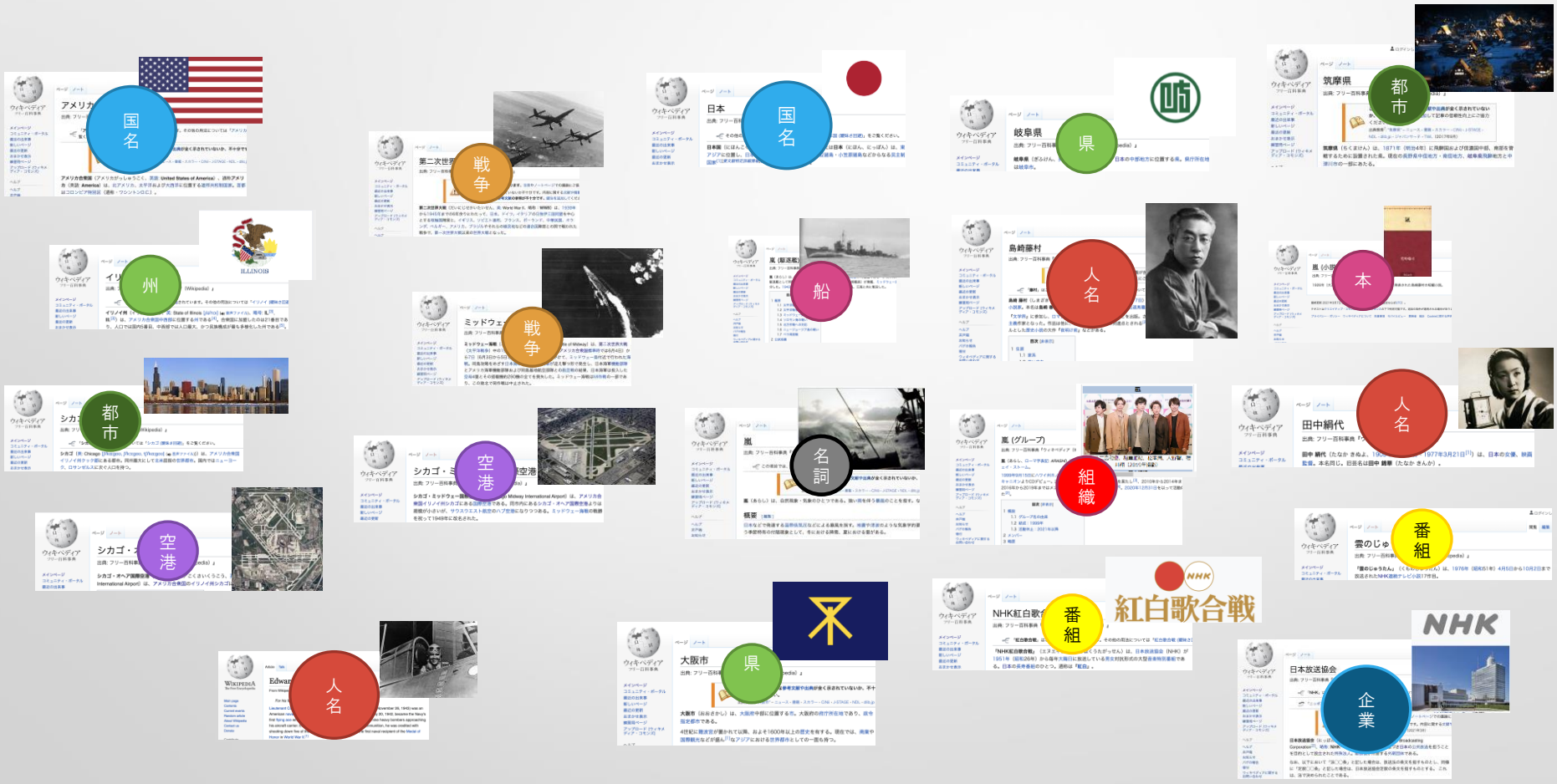
抽出した属性値を該当するWikipediaページに紐付け
（「嵐」はWikipediaページの「嵐（作品）」のこと！）

Wikipediaページ

（日本語は100万、英語は600万）



ステップ１（分類）

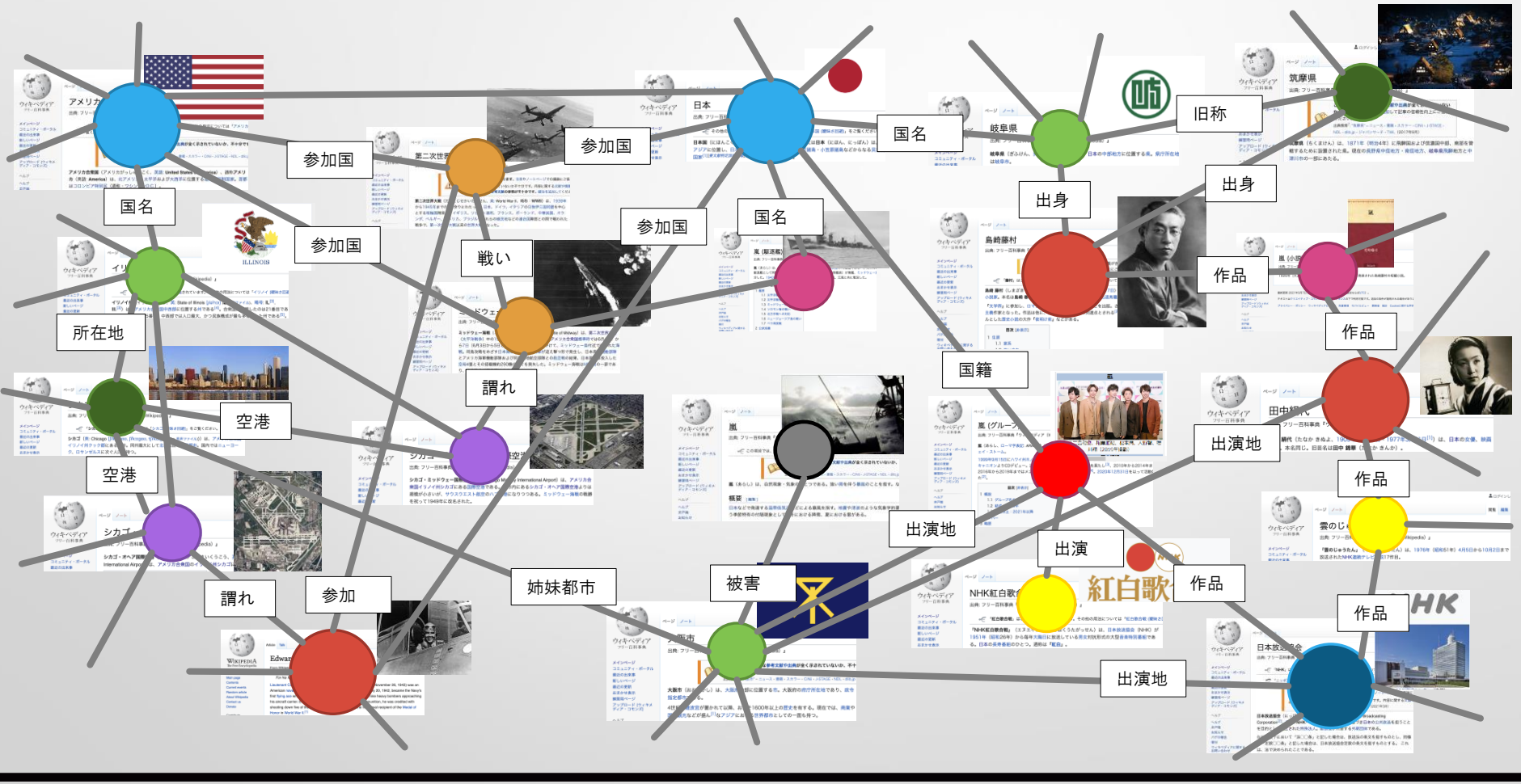


ステップ２
（属性値抽出）

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

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

WSの今後

- ２日目（８月１０日）13時

  Doorkeeper: https://c5dc59ed9782138303355fc8978.doorkeeper.jp/events/141143
  - 質疑応答
    - Slackで挙がった質問なども含む
  - ベースラインを超えたシステムの紹介
  - 精度向上のためのヒントなど共有
  - BERTによる他の自然言語処理タスクの紹介

- その後：BERTを使った色々なタスク（予告）
  - 今回は「分類」をBERTで実現
  - 次回は「属性値抽出」をBERTで実現
  - 次次回は。。。

森羅チームメンバー募集

このWSのその先に

- 森羅のテーマに研究として取り組みたい学生を募集
  - 最大3名（だいたいB4~D1を対象）
  - チームにアルバイト、インターンとして参加
  - 期限はまず今年度末（相談の上、来年度の延長も可能）
  - 2回目（8月10日）に応募方法をアナウンス

本日の講師の中山功太さんも、2017年のインターン学生

  - 社会人の希望者があれば、応相談

リーダーボード作成等の門脇さんも、社会人参加者