「森羅プロジェクト２０１８」最終報告会

UNISYS

Wikipedia構造化手法ご報告
～機械読解タスクとして解く～

日本ユニシス株式会社 石井愛

Foresight in sight

石井愛（Ai Ishii)

日本ユニシス株式会社 総合技術研究所 研究員

- 2010年頃～ 検索エンジンの評価・案件適用
- 2014年　NTCIR 11 RITE-VAL （情報検索+含意関係認識タスク）
- 2015年　NTCIR 12 QA-Lab2、「ロボットは東大に入れるか」プロジェクトにてセンター試験世界史を担当

■ 森羅プロジェクト参加のモチベーション

- 機械読解を情報検索に応用する研究に興味があり、日本語のデータで実験してみたかった

アジェンダ

- Wikipedia構造化手法
  - 全体の処理の流れ
  - 主な処理とその詳細
  - 実行環境・実行時間
- 結果と考察
  - 各カテゴリの学習・評価データでの結果
  - 考察
    - 参考）改善版の結果

Wikipedia構造化手法

解法の概要

- 機械読解タスク*として解く *ドキュメントを読んで質問に回答するタスク
  - 抽出する属性名を質問、対象ページをドキュメント、属性値を回答として学習し、未知のページの属性値を予測する

質問

小松飛行場 の ふりがなは？

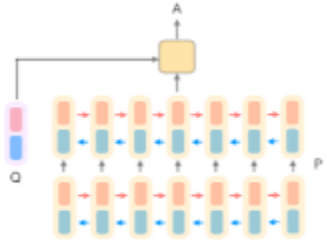
Wikipediaタイトル **属性名**

ドキュメント（対象のページのテキスト）

機械読解

回答（属性値）

こまつひこうじょう



Wikipedia構造化処理の流れ

1. 機械読解用データ作成（SQuAD形式に変換）

2. 機械読解（DrQA）による学習・予測

3. 予測結果をルールで補正

1．機械読解用データ作成（1/2）

- 前処理
  - ドキュメントデータの準備
    - Wikipedia本文とInfobox部分のテキストを、それぞれDBに格納
      - 本文：Cirrus Search ダンプデータのプレーンテキスト
      - Infobox部分：Wikipedia HTML データからXpathで抽出
        - » 項目名と値が両方抽出できた場合は、”[項目名]は[値]。”というテキストに変換
  - 単語ベクトル生成
    - FastTextを用いて日本語Wikipediaの本文をMecabで分割したデータから300次元の単語ベクトルを生成
      - FastText: https://github.com/facebookresearch/fastText
      - Mecab: http://taku910.github.io/mecab/

1．機械読解用データ作成（2/2）

- Wikipedia構造化学習データをSQuAD形式に変換
  - ドキュメント＝Wikipedia本文
  - 質問＝“[タイトル]の[属性名]は？”
  - 回答＝該当する属性値が最初に出てきた箇所
    - 開始位置が必要だったため、初出の個所が一番説明されているだろうと仮定
    - 属性値がない場合、ドキュメントの最初に記号”φ”を付加し、回答として設定

機械読解のデータセットSQuADの例

- Stanford Question Answering Dataset (SQuAD) [Rajpurkar 16]の例
  - ドキュメント、質問、回答の組が与えられる
  - 回答として、回答文字列とドキュメント内の開始位置が与えられ、ドキュメント内のテキストスパンを選択する

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

ドキュメント

What causes precipitation to fall?
gravity start=18

質問①
回答①

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel start=35

質問②
回答②

図は論文より

変換データの例

Wikipedia構造化データ形式


```
{
 "ENE":"空港名",
 "Name":"小松飛行場",
 "WikipediaID":100192,
 "Attributes":{   "ふりがな":["こまつひこうじょう"],
          "IATA（空港コード）":["KMQ"],
          "ICAO（空港コード）":["RJNK"],
          "別名":["Komatsu Airbase","Komatsu Airport","小松空港","FAC4017小松補助飛行場"],
          "名称由来":[],
          "名称由来人物の地位職業名":[],
          "国":["日本"],
          "年間利用客数":[],
"年間利用者数データの年":[],
          "年間発着回数":[],
          "年間発着回数データの年":[],
          "座標・経度":["東経136度24分27秒","東経136.40750度"],
          "座標・緯度":["北緯36度23分38秒","北緯36.39389度"],
          "所在地":["石川県小松市","むじなが浜"],
          "旧称":[],
          "標高":["6 m","18 ft"],
```


SQuADデータ形式

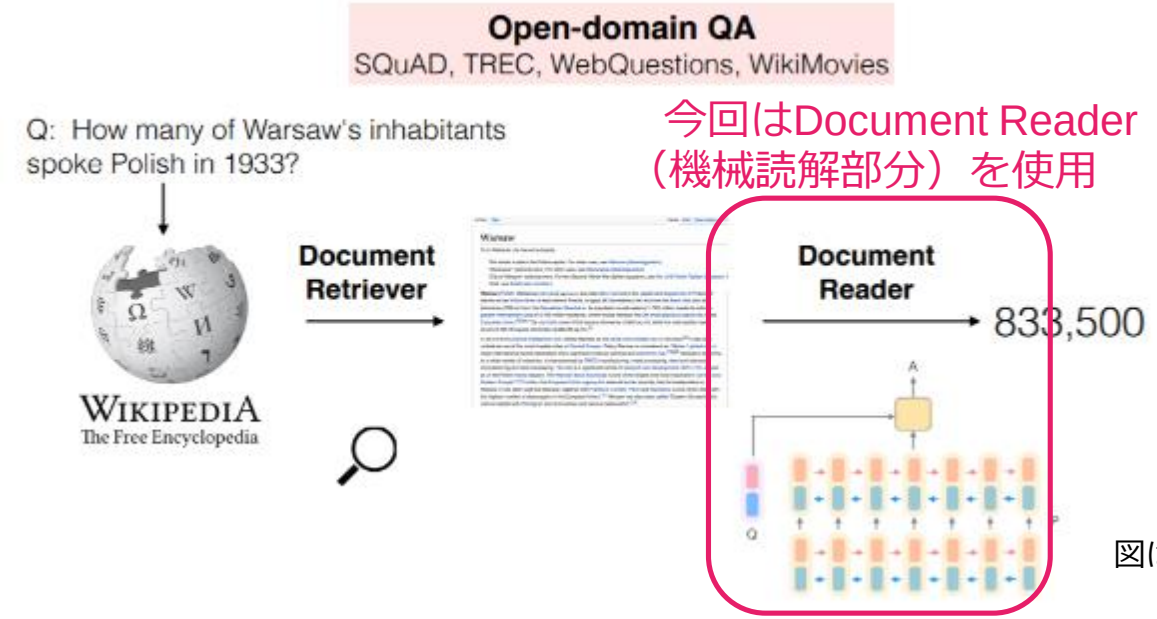

```
{
 "ENE":"空港名",
 "WikipediaID":100192，
 "title":"小松飛行場",
 "paragraphs":[ {
      "context":"φ小松飛行場（小松空港）、Komatsu Airbase (Komatsu Airport)。ターミナルビル。IATA: KMQ - ICAO: RJNK。...小松飛行場(こまつひこうじょう)は、石川県小松市にある共用飛行場である。...",
      "qas":[
       {
         "id":"100192_1“,
         "question":"小松飛行場のふりがなは？“,
         "answers":[{
              "answer_start":661,
              "text":"こまつひこうじょう"}]
       },
         …
       {
          "id":"100192_5“,
          "question":"小松飛行場の名称由来は？“,
          "answers":[{
               "answer_start":0,
               "text":"φ"
           }]
        },
```

2. 機械読解による学習・予測（1/2）

- 機械読解の実装
  - 機械読解部分は公開されているDrQAの実装を利用
  - Mecabを用いた日本語用のトークナイザを追加

DrQAの概要（1/2）

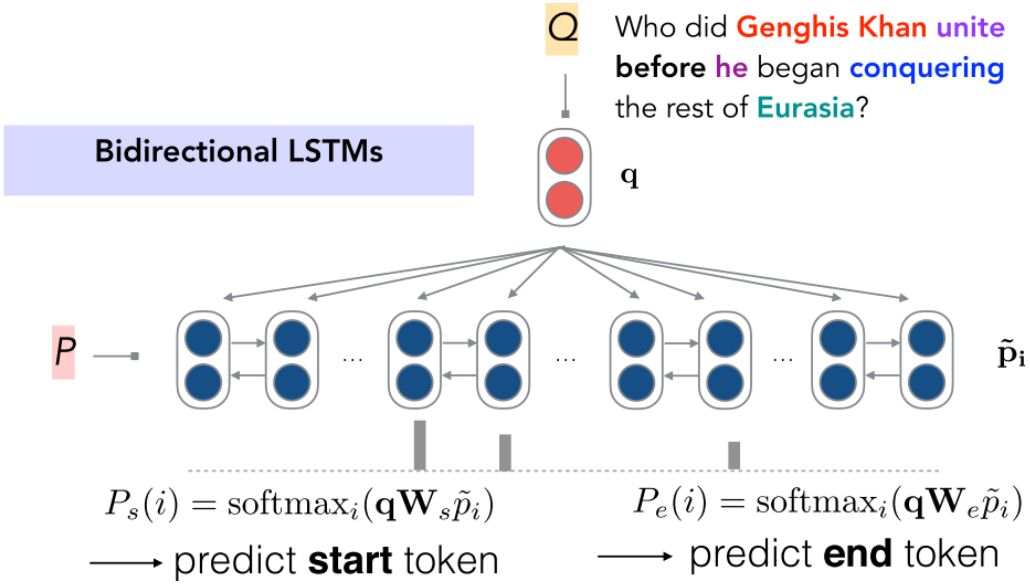
- Reading Wikipedia to Answer Open-Domain Questions [Chen, 2017] の実装
  - https://github.com/facebookresearch/DrQA
- 機械読解タスクではあらかじめ与えられるドキュメントを、Wikipediaから検索して取得する処理を加えて、オープンドメインな質問応答を実現



図は論文より

DrQAの概要（2/2）

- Document Readerの仕組み
  - 学習済みの単語分散表現・品詞などの単語特徴・質問中の単語との一致などを入力としてRNNを使用し、回答のstartとendの位置を予測する
    - 品詞、単語頻度、質問中の単語との一致、単語ベクトルの類似度
  - SQuADのリーダーボードのトップの成績とほぼ変わらない精度という報告（2017年2月時点）



図は著者のポスターより

2. 機械読解による学習・予測（2/2）

- 学習
  - 全カテゴリのデータを用いて１つのモデルを作成
    - ５カテゴリ×600ページ×属性数の72000件
      - 95%を訓練データ、5%を開発データに用いた
- 予測
  - TOP１０の答えとスコアを出力

3. 予測結果をルールで補正

- ルールで予測結果を補正
  - 十数個ルールを作成
    - 密度、温度などの数値データを正規表現でチェック
    - ふりがながカタカナとひらがなのみかをチェック
    - 別名、正式名称からふりがなを除外
    - 居住地、所在地から国を除外
    - 没年月日と生年月日が同じ場合は削除し、死亡関連項目の値を削除
  - 答えが1つになりやすい属性はTOP 1 に絞り、複数になりやすい場合は足切りするスコアを小さく設定

- 後日最終的な実装による結果からルールによる補正を抜いてみた結果
  - Presitionが0.06～0.1下がったもののrecallが0.01～0.06上がり、差分はf1の平均値で0.01程度（学習データ）

実行環境・時間

開発・実行環境

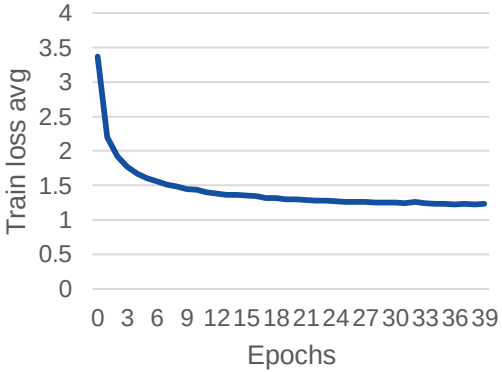
- GPU搭載サーバ1台
  - インテル(R) Core(TM) i7-7700 プロセッサー　4 コア / 8 スレッド / 3.60GHz
  - 64GB メモリ
  - NVIDIA GeForce GTX1080Ti / 11GB

実行時間

- 学習　40 epoch　約50時間
- 予測
  - 企業名　26799件
  - 空港名　1465件
  - 化合物名　3975件

-- 以下は期日後に提出 --

- 人物名　249023件　約72時間
- 市区町村名　45713件　約12時間



結果と考察

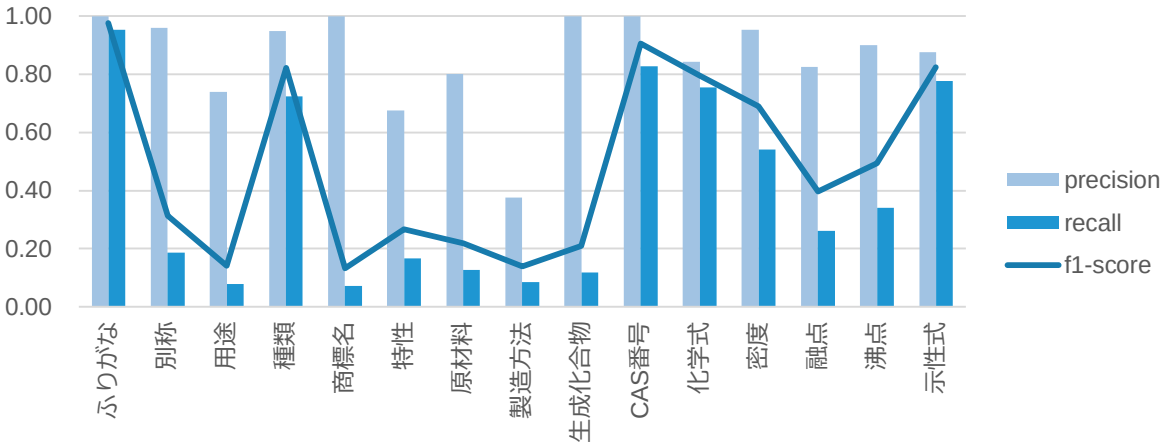
カテゴリ毎の結果　化合物名

学習

(N=1282)

Avg all

Prec	0.82
Recall	0.27
F1	0.36

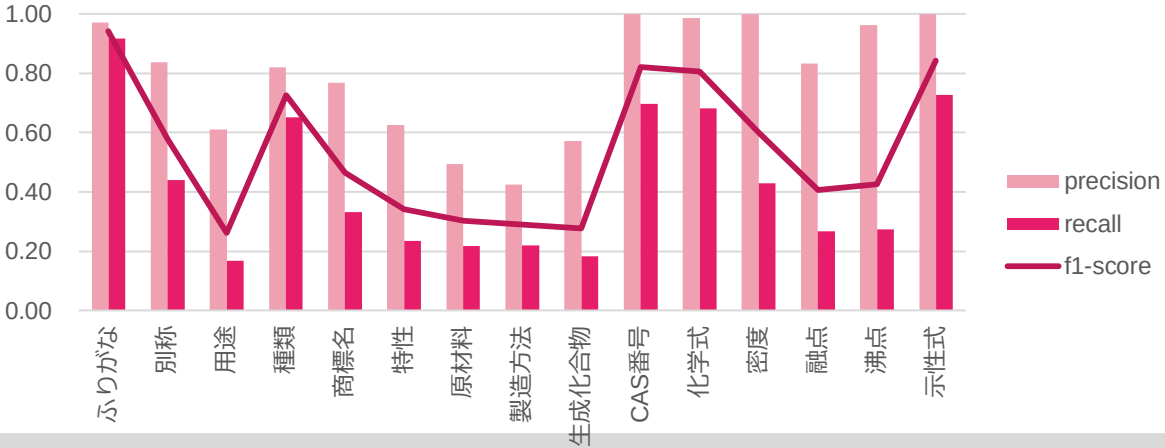


評価

(N=2181)

Avg all

Prec	0.72
Recall	0.34
F1	0.45

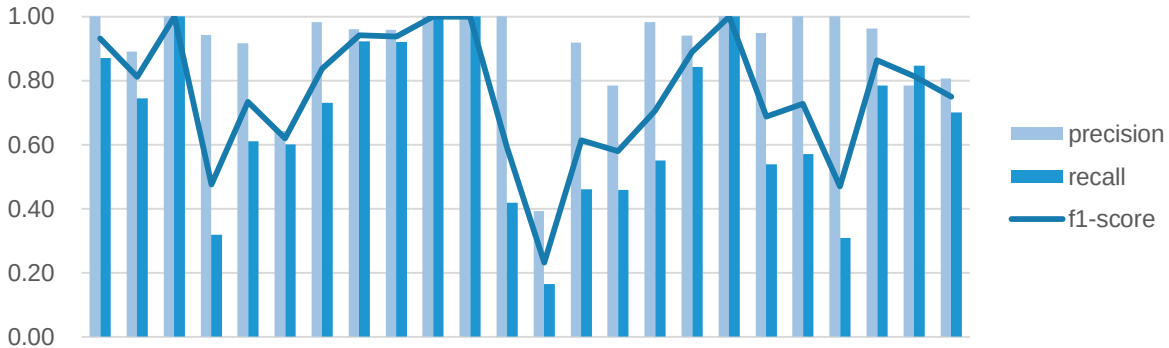


カテゴリ毎の結果　空港名

学習

(N=1273)

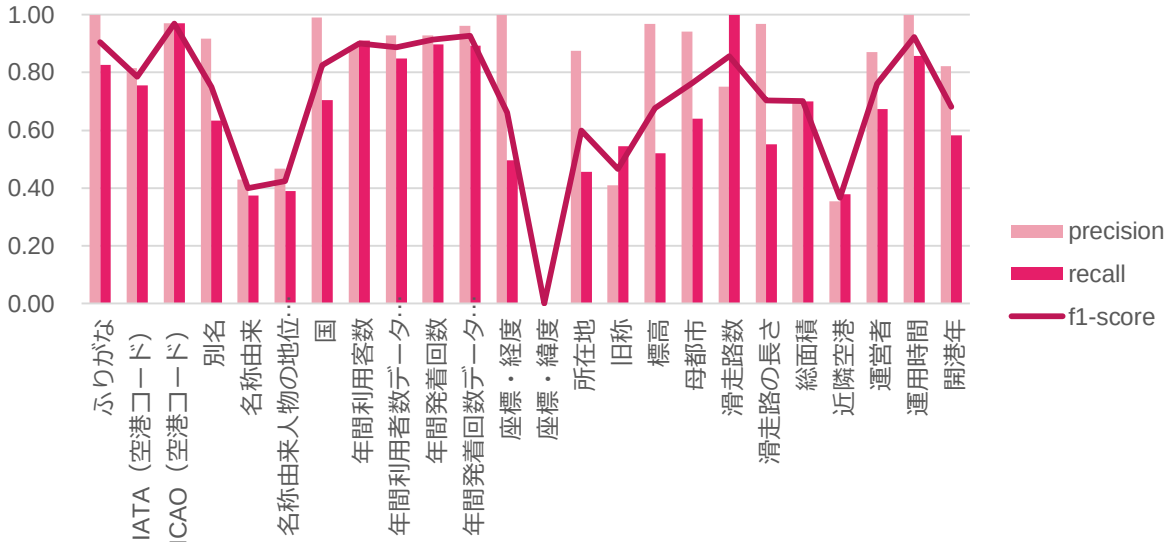
Avg all
Prec 0.89
Recall 0.56
F1 0.66



評価

(N=2061)

Avg all
Prec 0.82
Recall 0.57
F1 0.66



UNISYS

カテゴリ毎の結果　企業名

学習

(N=1927)

Avg all

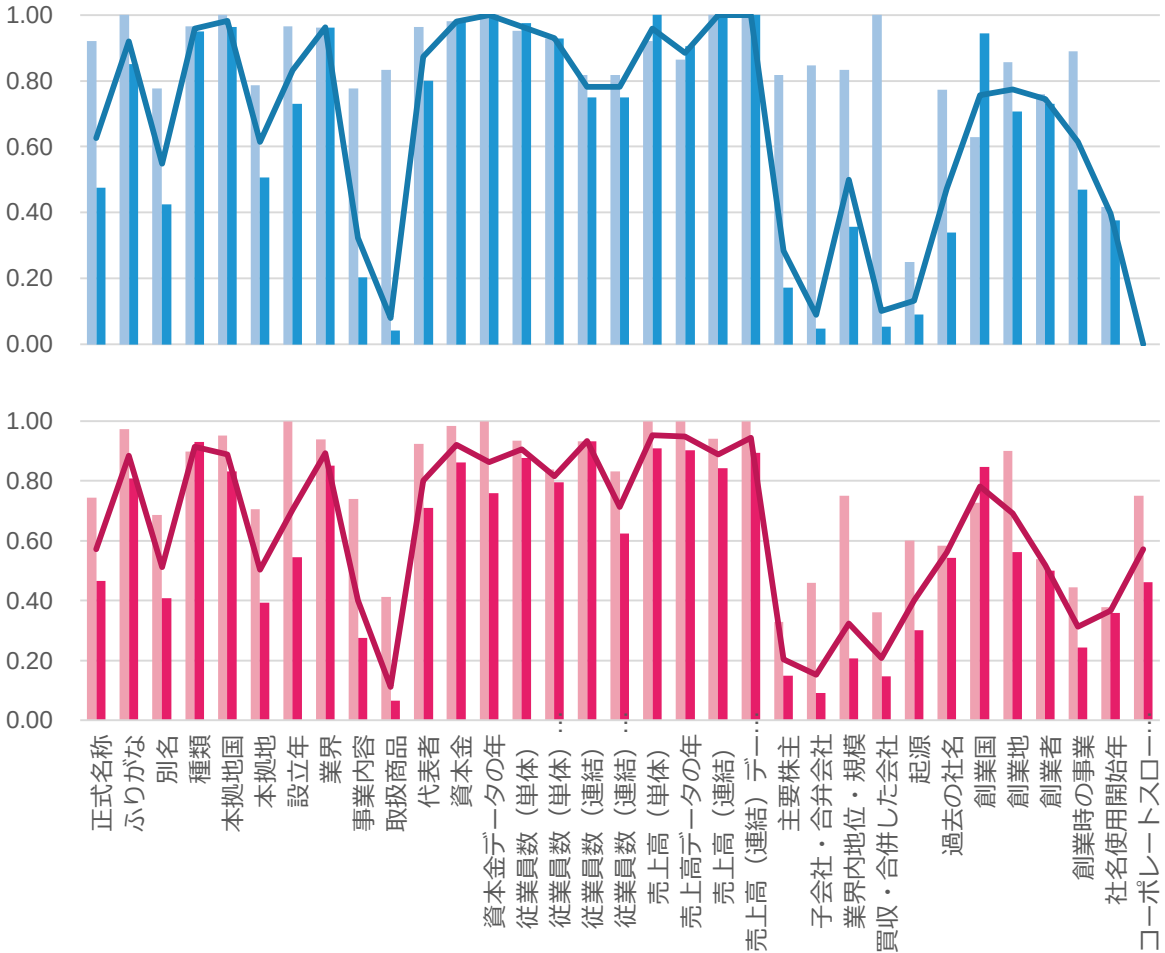
Prec	0.85
Recall	0.44
F1	0.50

評価

(N=3066)

Avg all

Prec	0.67
Recall	0.42
F1	0.49



カテゴリ毎の結果　市区町村名 (参考)

学習

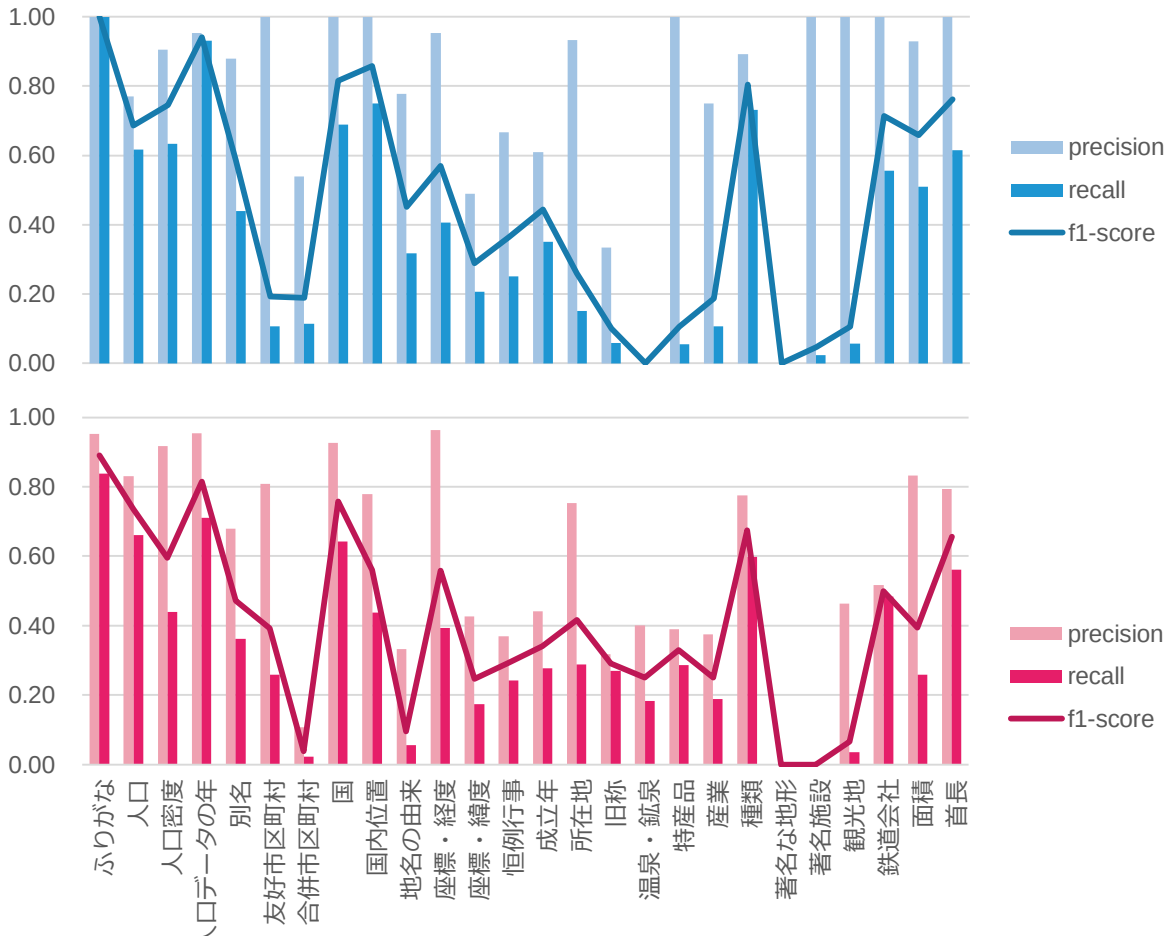
(N=1269)

Avg all
Prec 0.85
Recall 0.35
F1 0.44

評価

(N=2699)

Avg all
Prec 0.62
Recall 0.30
F1 0.38



カテゴリ毎の結果　人名 (参考)

学習

(N=1343)

Avg all

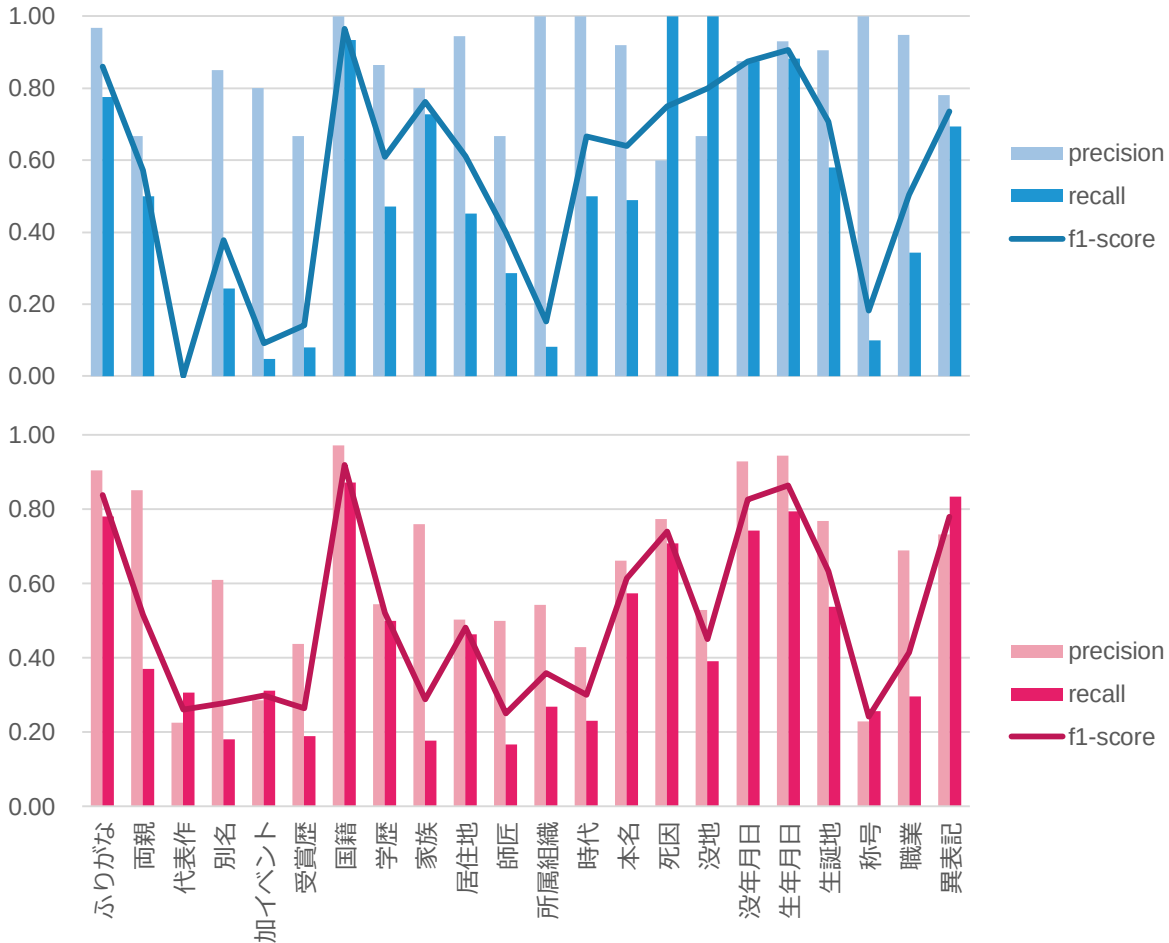
Prec	0.88
Recall	0.32
F1	0.41

評価

(N=2839)

Avg all

Prec	0.59
Recall	0.36
F1	0.43



考察

- 機械読解の１つのモデルで全体的にある程度の精度で属性抽出ができた
- 属性値が１つの属性については精度が高い

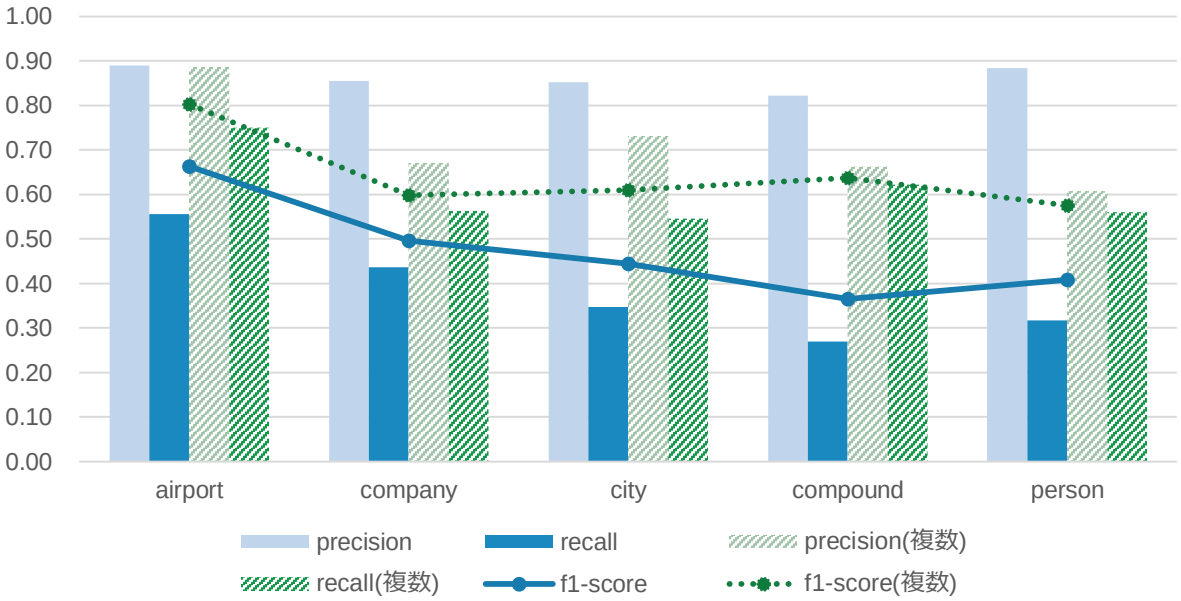
- 複数の値となる属性抽出に弱い
  - DrQAの実装が複数回答からの学習に対応していないため

→複数回答対応版

複数回答に対応するため、[鈴木, 2018]を参考にネットワークの最終層を各トークンが回答となる確率を算出するように変更

参考）複数回答対応版

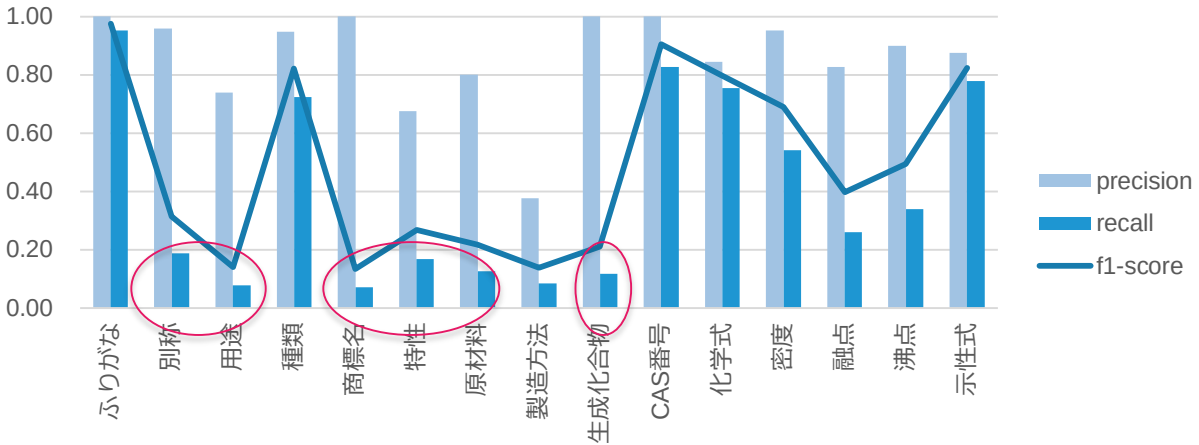
- 提出した実装と複数回答対応の実装を各カテゴリの属性の平均で比較
  - 青系　提出した実装
  - 緑系　複数回答対応版
- Presitionは0.0～0.28悪化したものの、Recallは0.13~0.35改善し、F1も0.10~0.27改善しすべてのカテゴリで0.5を超えた



参考）複数回答対応版との比較（化合物名）

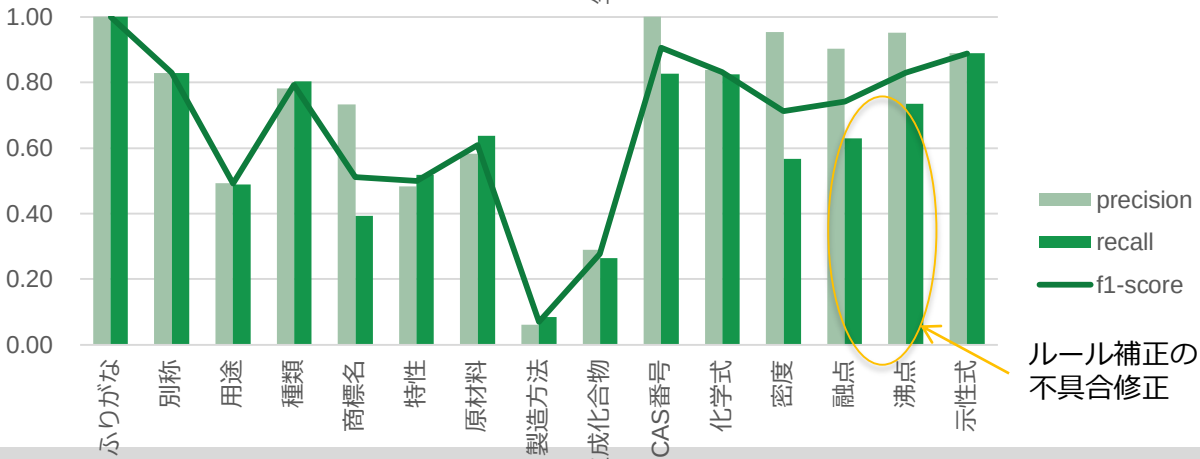
提出版
（学習）
(N=1282)

Avg all
Prec 0.82
Recall 0.27
F1 0.36



複数回答
対応版
（学習）
(N=1282)

Avg all
Prec 0.66
Recall 0.62
F1 0.64



ルール補正の
不具合修正

謝辞

- 森羅プロジェクトのみなさま

  貴重な日本語のデータを使用して実験できる良い機会となり感謝しております

- SanSan株式会社　奥田様 高橋様

  データ確認用のウェブアプリ、評価スクリプトともに重宝させていただきました

- 会場のみなさま

  ご清聴どうもありがとうございました

参考文献

- [Chen, 2017] Chen, D., Fisch, A., Weston, J., and Bordes, A.: Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers), pp. 1870–1879 (2017)
- [Rajpurkar, 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.2383-2392 (2016)
- [鈴木, 2018] 鈴木 正敏, 松田 耕史, 岡崎 直観, 乾 健太郎: 読解による解答可能性を付与した質問応答データセットの構築, 第24回言語処理学会年次大会, pp. 702-705 (2018)

Foresight in sight

UNISYS