sansan

森羅：Wikipedia構造化プロジェクト2018

Sansan最終報告

Sansan株式会社

高橋寛治, 奥田裕樹

Sansanのミッション

ビジネスの出会いを資産に変え、
働き方を革新する

Creating a resource from everyday business
encounters and transforming
the way the world works.

研究開発部門の位置づけ

Sansan株式会社

Sansan事業部

法人向け名刺管理サービス
Sansanの開発、提供

Eight事業部

個人向け名刺アプリ
Eightの開発、提供

Data Strategy & Operation Center

データ統括部門

R&D

データ分析・研究開発
（画像処理／機械学習・AI）

名刺データ化における研究開発のビジネスへの貢献

数億枚の名刺を精度99.9%でデータ化

名刺1枚あたりの入力コストを約80%削減

名刺画像の読取

×

画像処理・言語処理など
情報処理技術

名刺情報の入力

sansan

名刺DB

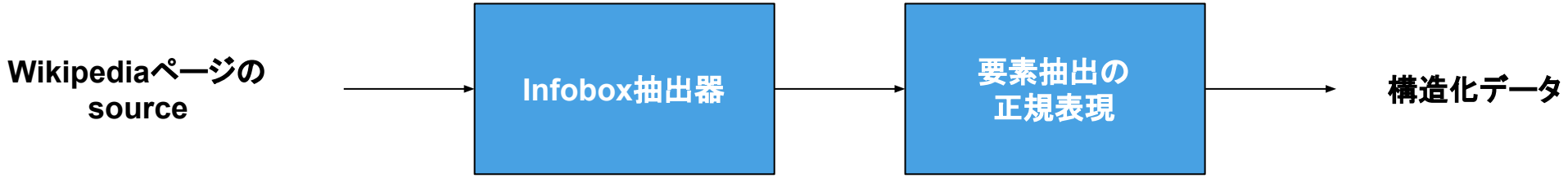
sansan

一枚まとめ

- 目的
  - 企業情報を抽出して自社ビジネスに活用したく、技術領域が重なり参加
- 手法
  - 短期間での抽出を検討するために、ルールベースと固有表現抽出を採用
- 結果
  - カテゴリー:Companyに対して抽出を実施
- その他
  - 評価用スクリプトや閲覧ツールの共有
    - > https://gist.github.com/kanjirz50/616b3a1c069dc4b0a4d9357457f6a105
    - > http://shinra-1027558540.ap-northeast-1.elb.amazonaws.com/

手法概要

Infoboxを対象にした正規表現による素朴なルールベース



- Infoboxからの要素抽出の概略

1.1. カテゴリに分類し、カテゴリ名を正規化

1.2. ノイズを削除し、カテゴリに対応する正規表現で属性抽出

infoboxパースと要素抽出の例

```
def _extract_sales(self, category_items):
    _ = " ".join([item for i, item in enumerate(category_items['売上高'])])
    if "連結" in _:
        # 売上高(連結)
        matched = re.search(r"[\d,]+ ?(million) ?(us\$)", _, re.IGNORECASE)
        if matched:
            category_items['売上高(連結)'] = [matched.group(0)]

        # 売上高(連結)データの年
        matched = re.search(r"\d+年\d月[末期]?", _)
        if matched:
            category_items['売上高(連結)データの年'] = [matched.group(0)]

def _extract_employees(self, category_items):
    _ = " ".join([item for i, item in enumerate(category_items['従業員数'])])
    # 従業員数(単体)
    matched = re.search(r"[\d,]+ ?(人)", _, re.IGNORECASE)
    if matched:
        category_items['従業員数(単体)'] = [matched.group(0)]

    # 従業員数(単体)データの年
    matched = re.search(r"\d+年\d月[末期]?", _)
    if matched:
        category_items['従業員数(単体)データの年'] = [matched.group(0)]
```

- 基本的にはデータを見て対応関係やルールを策定
  - 正規表現ベース
- 特定の項目は表記の多様性が高い
  - 売上や従業員数、資本金など、 年度や時期で変動する項目

結果

	attribute	precision	recall	f1-score	support
0	正式名称	0.949825	0.698113	0.804745	1166
1	ふりがな	0.000000	0.000000	0.000000	325
2	別名	0.000000	0.000000	0.000000	334
3	種類	0.982103	0.783929	0.871897	560
4	本拠地国	0.966408	0.653846	0.779979	572
5	本拠地	0.538211	0.399758	0.458766	828
6	設立年	0.329094	0.255871	0.287900	809
7	業界	0.538824	0.416364	0.469744	550
8	事業内容	0.640074	0.165865	0.263459	2080
9	取扱商品	0.000000	0.000000	0.000000	2563
10	代表者	0.308943	0.243200	0.272158	625
11	資本金	0.715278	0.656051	0.684385	471
12	資本金データの年	0.298969	0.188312	0.231076	154
13	従業員数(単体)	0.639706	0.240331	0.349398	362
14	従業員数(単体)データの年	0.366279	0.258197	0.302885	244
15	従業員数(連結)	0.000000	0.000000	0.000000	101
16	従業員数(連結)データの年	0.000000	0.000000	0.000000	99

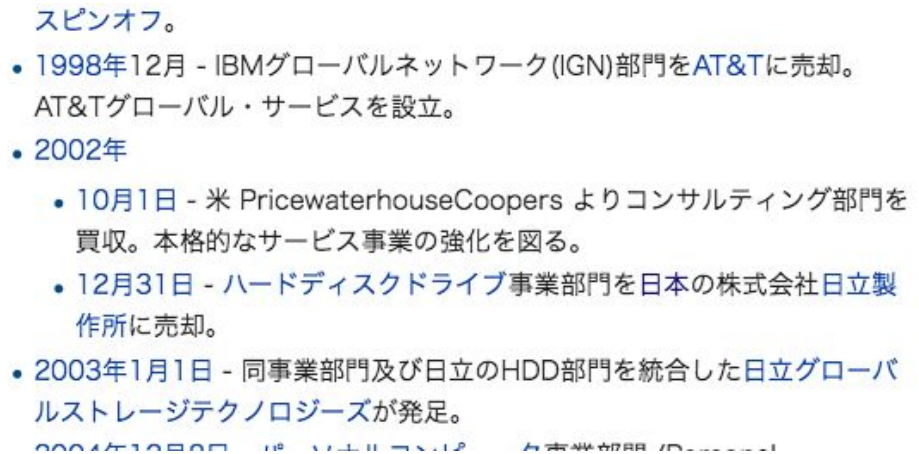
	attribute	precision	recall	f1-score	support
17	売上高(単体)	0.000000	0.000000	0.000000	240
18	売上高データの年	0.000000	0.000000	0.000000	218
19	売上高(連結)	1.000000	0.024590	0.048000	122
20	売上高(連結)データの年	0.986667	0.627119	0.766839	118
21	主要株主	0.283626	0.174932	0.216397	1109
22	子会社・合弁会社	0.777397	0.124247	0.214252	1827
23	業界内地位・規模	0.000000	0.000000	0.000000	148
24	買収・合併した会社	0.000000	0.000000	0.000000	621
25	起源	0.000000	0.000000	0.000000	157
26	過去の社名	0.000000	0.000000	0.000000	504
27	創業国	0.000000	0.000000	0.000000	488
28	創業地	0.000000	0.000000	0.000000	166
29	創業者	0.000000	0.000000	0.000000	236
30	創業時の事業	0.000000	0.000000	0.000000	210
31	社名使用開始年	0.000000	0.000000	0.000000	542
32	コーポレートスローガン	0.000000	0.000000	0.000000	74

本文からの情報抽出は検討で終了

- IOB2タグ形式で学習データを作成
  - タグの種類
    - 固有表現ではない単語はO
    - 抽出対象はBIでタグ付け
  - タグ付け方法
    - 本文一致箇所をすべてタグ付け
- 結果
  - 学習での文脈が少なく過学習気味で断念

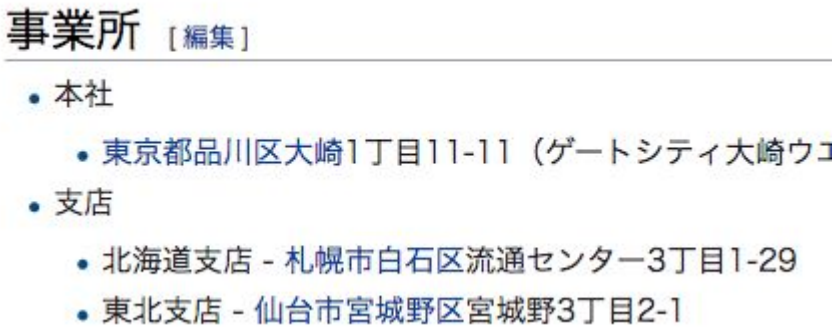
振り返り - 今後必要になりそうな技術

- テキスト上で半構造化された部分の適切なパース
  - 年表や箇条書き、書式の定まっていない表構造への対応が必要
    - > ルールベースで愚直にパーサーを実装する、文章中の距離的な近さから対応関係を獲得する, etc…
    - > 画面上のレイアウトや位置関係を文章の解釈に利用する必要がある



スピンオフ。
- 1998年12月 - IBMグローバルネットワーク(IGN)部門をAT&Tに売却。AT&Tグローバル・サービスを設立。
- 2002年
  - 10月1日 - 米 PricewaterhouseCoopers よりコンサルティング部門を買収。本格的なサービス事業の強化を図る。
  - 12月31日 - ハードディスクドライブ事業部門を日本の株式会社日立製作所に売却。
- 2003年1月1日 - 同事業部門及び日立のHDD部門を統合した日立グローバルストレージテクノロジーズが発足。

https://ja.wikipedia.org/wiki/IBM



事業所 [編集]
- 本社
  - 東京都品川区大崎1丁目11-11（ゲートシティ大崎ウ
- 支店
  - 北海道支店 - 札幌市白石区流通センター3丁目1-29
  - 東北支店 - 仙台市宮城野区宮城野3丁目2-1

https://ja.wikipedia.org/wiki/日本石油輸送

振り返り - 会社の持つ資産の活用

- Sansanの持つ会社情報の資産を活かしきれなかった
  - 会社名などをデータとして保有していても固有表現抽出に活かすのが難しい
    - 所有辞書上の表記とwikipedia文中の表記の違い
    - （弊社の場合は）文章としてコーパス等を保有しているわけではない
  - 抽出ロジック全体のシステムのどこかには利用できそうではあった
    - 例）抽出した文字列が会社名らしいかどうかの2値分類の判定器、など
    - タスクをもっと分解する必要がある

- オープンイノベーションとして企業情報を公開していく意義
  - むやみに公開してもそれを活かす道がないと仕方がない
    - 公開するデータの種類、データ形式、具体的な活用事例等

振り返り - 自由に使える日本語コーパスや固有表現抽出器

- 手軽に使えるリソースが限られている
  - 企業にいるとBCCWJや新聞コーパス等が自由に利用できない
    - 買えばいいって話なんですが……
  - まずはwikipediaなどからコーパスを作り……固有表現抽出器を実装して……
    - ベースラインにたどり着くまでにかなりの時間と労力を要する
- とりあえず何かしら動くものが簡単に使えるという状態が理想
  - 研究者でもエンジニアでも、日本語の形態素解析で躓く人はいない
    - MeCab、CaboCha、SudachiなどのOSS、独自拡張ができる仕組み
  - 固有表現抽出において、自由な(拡張性のある)固有表現抽出器があると嬉しい
    - MeCabにおける、NEologdや独自辞書のような関係
    - 抽出する固有表現のドメイン、精度等のハードルはあるものの

まとめ

- ルールベースおよび固有表現抽出を利用した抽出
  - カテゴリー:Companyを対象に実施
  - 精度面では、あまり貢献できませんでした……
- 感想
  - Wikipediaのデータを用いた研究開発の知見を蓄積できた点は良かったです
  - 企業としてシェアタスクへの取り組みを考えるきっかけになりました
    > 企業側からのアプローチや、情報公開の方法を模索するトライアルとして
  - 他の企業や研究室での取り組み方など、ぜひ情報交換させていただければと思います

ビジネスの出会いを
資産に変え、
働き方を革新する

Creating a resource from everyday business encounters and transforming the way the world works.


sansan