Cappuccino：構造データに対するブートストラッピング手法

花房 諒[1]　小山田 昌史[2]

1：関西学院大学 大学院 理工学研究科　2：日本電気株式会社 データサイエンス研究所

概要

構造データからの関係抽出タスクを扱い，構造データに対するブートストラッピング手法である Cappuccino を提案する．

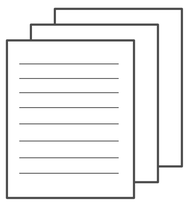
構造データからの関係抽出においては，
文字列パターンだけを用いる
従来のブートストラッピング手法よりも
データがもつ構造を利用する提案手法が
高い性能になることを確認する．

背景

関係抽出

例：上位下位関係

入力：文書（非構造データ）



出力：インスタンス獲得

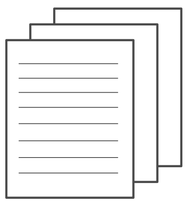
(animal, dog)
(fruit, apple)
(country, Japan)
⋮

We present *Cappuccino*, a bootstrapping method for structural data, which extracts target pairs of words by treating structures in data as patterns that connect pairs of words. ･･･

人手による作業は
コストが高い

ブートストラッピング手法

入力：シードインスタンス
(animal, cat)
⋮
文書



繰り返し
- パターン抽出
  - 信頼度によるランキング
- インスタンス抽出
  - 信頼度によるランキング

出力：インスタンス獲得

Espresso [Pantel+, COLING'06]：代表的な手法

パターンの信頼度：
$$r_\pi(p) = \frac{1}{|I|}\sum_{i\in I}\frac{pmi(i,p)}{\max pmi} r_\iota(i)$$

インスタンスの信頼度：
$$r_\iota(i) = \frac{1}{|P|}\sum_{p\in P}\frac{pmi(i,p)}{\max pmi} r_\pi(p)$$

$$pmi(i,p) = \log_2 \frac{|i,p|}{|i,*||*,p|}$$

自己相互情報量の推定量

I と P：各繰り返し時の
それぞれの集合

$|i,p|$：文単位での共起回数

構造データを対象とするブートストラッピング手法を提案

構造の例：　テーブルやリスト，カテゴリ名，リンク
構造データの例：　HTMLやプレゼンテーションスライド

* 構造データに対する関係抽出手法である
Founduer [Wu+, arXiv'17] は構造パターンが事前に必要

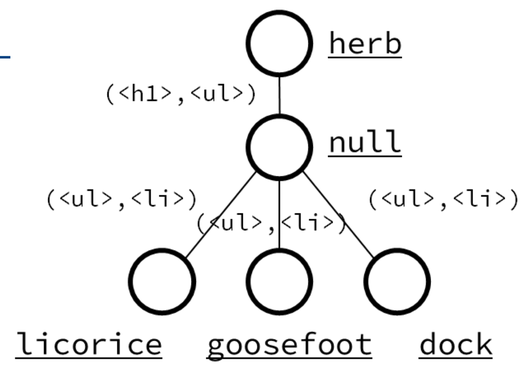
Cappuccino（提案手法）

Espresso の信頼度関数を利用：　共起回数が必要

構造データをグラフへ変換

```
<h1>herb</h1>
  <ul>
    <li>licorice</li>
    <li>goosefoot</li>
    <li>dock</li>
  </ul>
```

HTML の例

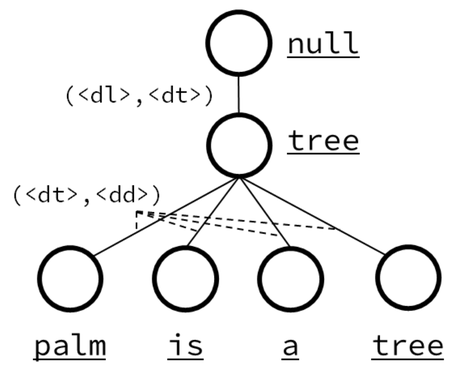


ノードラベル：　タグで囲まれる単語
エッジラベル：　親と子のタグの組

通常の文を含む場合

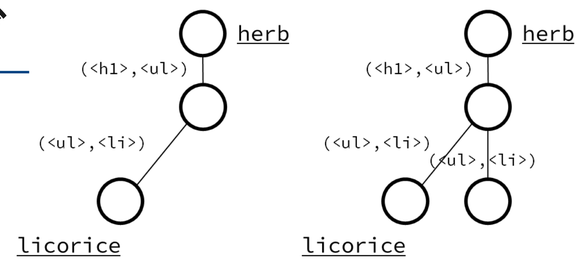
```
<dl>
  <dt>tree</dt>
  <dd>palm is a tree.</dd>
</dl>
```

タグに囲まれた文は分解



頻出サブグラフマイニング

包含関係にあるサブグラフは
パターンとインスタンスの
別の組み合わせとして扱う



インスタンスとパターンとの共起回数をカウント可能

実験

人工データ

WordNet から植物 100 種類

正しい上位下位関係をもつインスタンス 100 個
を含む実験のために作成した HTML ファイル

インスタンス 10 個ごとに，それらを繋いでいる
構造パターンと文字列パターンの個数と比率が異なる

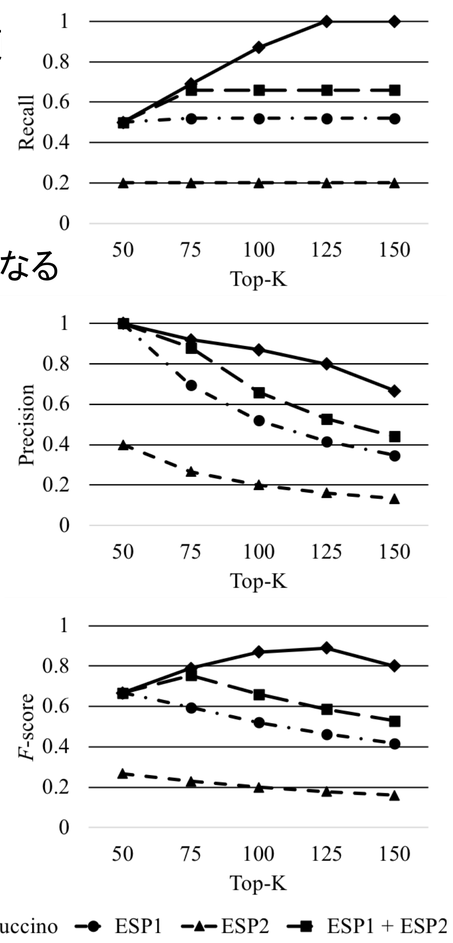
比較手法

3種類の Espresso

1. ESP1：通常の文を扱うためにタグで
   囲まれた部分を文とする．
2. ESP2：単語の組を扱うためにタグが間に
   ある部分を文とする．
3. ESP1+ESP2：1 と 2 を合わせたものを
   文とする．

パラメータ

シードインスタンス数：2
繰り返し回数：100
サブグラフのノードの最大数：4
少なくとも1 回出現するグラフをカウント



結論

構造データに対する関係抽出のタスクにおいて，
既存手法である Espresso と比較し，
Cappuccino が再現率と精度，F 値の
すべてで高い性能となった．

今後の展望

属性やメタ情報を利用することにより，
エッジラベルの種類を増加させ，実データに適用する．