名詞カテゴリからの関係知識獲得に向けて

- カテゴリ間の関係を獲得
  - 同じ意味の表現をまとめあげ，カテゴリ間の関係とする

パナソニックが三洋を子会社化する
山崎製パンが不二家を完全子会社化する
任天堂がセガを買収する
} <企業>が<企業>を*子会社化する*

- なぜ関係知識が欲しいのか？
  - 因果関係の獲得や推論に必要

<企業>が<企業>に資本業務提供する
<企業>が<企業>に資本提供する
→
<企業>が<企業>を子会社化する
<企業>が<企業>を買収する

  - 人手で構築することは大変
- 既存手法を日本語に適用する

東北大学
高瀬翔 岡崎直観 乾健太郎

関連研究

- 推論規則の獲得[Lin and Pantel 01]
  - 動詞の項を元に相関のあるパターンを獲得

    Dickens is author of Oliver Twist
    Dickens wrote Oliver Twist → X is author of Y = X wrote Y
- カテゴリ間の関係を獲得[Mohamed+ 11, Nakashole+ 12]
  - カテゴリ間のインスタンス対を元にクラスタリングを行い関係を獲得
- 動詞間の関係の推定[chklovski and Pantel 04]

to X and then Y → XとYの間には因果関係がある

モチベーション

- [Mohamed+ 11]の日本語への適用を行う
- 因果関係が同じクラスタに属してしまう問題に対処
  - パターン間の文中での共起頻度から同義関係を特定

全体のながれ

企業：パナソニック，三洋，Apple，…
製品：Let's note，エネループ，Mac，…
…

パターンの抽出：
カテゴリ対のインスタンスを含む
パターンを抽出

コーパス

企業，企業：
XがYを子会社化する
XがYを完全子会社化する
XがYに勝利した
…

パターンのクラスタリング：
カテゴリ対毎に抽出したパターンをクラスタリング
因果関係にあるパターンを同一のクラスタに入れない

XがYを子会社化する
XがYを完全子会社化する
XがYを買収する

<企業>が<企業>を子会社化する

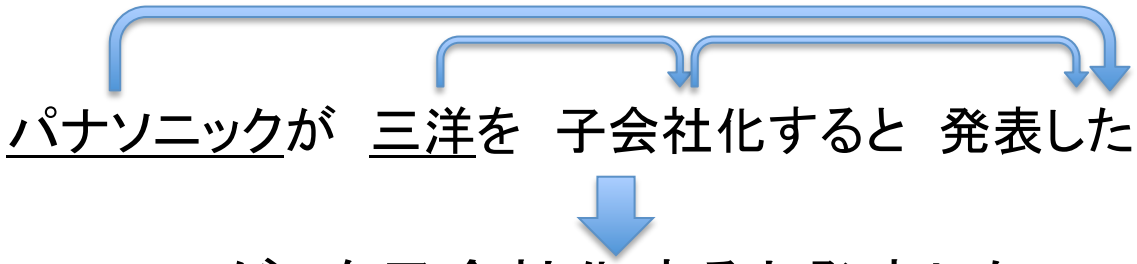
同義と言い換え表現
のみからなるクラスタ

XがYに勝利した
XはYを圧倒した

<企業>が<企業>に勝つ

パターンの抽出

- パターン
  - [Mohamed+ 11]：2つのインスタンスの間にある文字列
    - Google buys Youtube → X buys Y
  - 今回：係り受け解析結果の上で2つのインスタンスを結ぶ最短パス
    - パス上にある単語を出現順に並べた単語列



- 制約
  - 動詞かサ変名詞を含む
  - 主語を起点とする
    - 「Xが」または「Xは」を主語と考える

クラスタリング

- 同義や言い換え関係にあるパターンのみからなるクラスタを作成
  - 因果関係や反意にある関係のパターンを同一クラスタに属させない
  - 同義や言い換え関係にあるパターン＝同一文中で出現しない
    - 同一文中での共起度co_occur(p_1, p_2)がしきい値より高いパターンをマージしない

$$co\_occur(p_1, p_2) = \frac{co(v_{p_1}, v_{p_2})}{freq(v_{p_1}) * freq(v_{p_2})}$$

v：動詞やサ変名詞
freq(v)：vの出現回数
co(v_{p1},v_{p2})：同一文中でのv_{p1}とv_{p2}の共起頻度

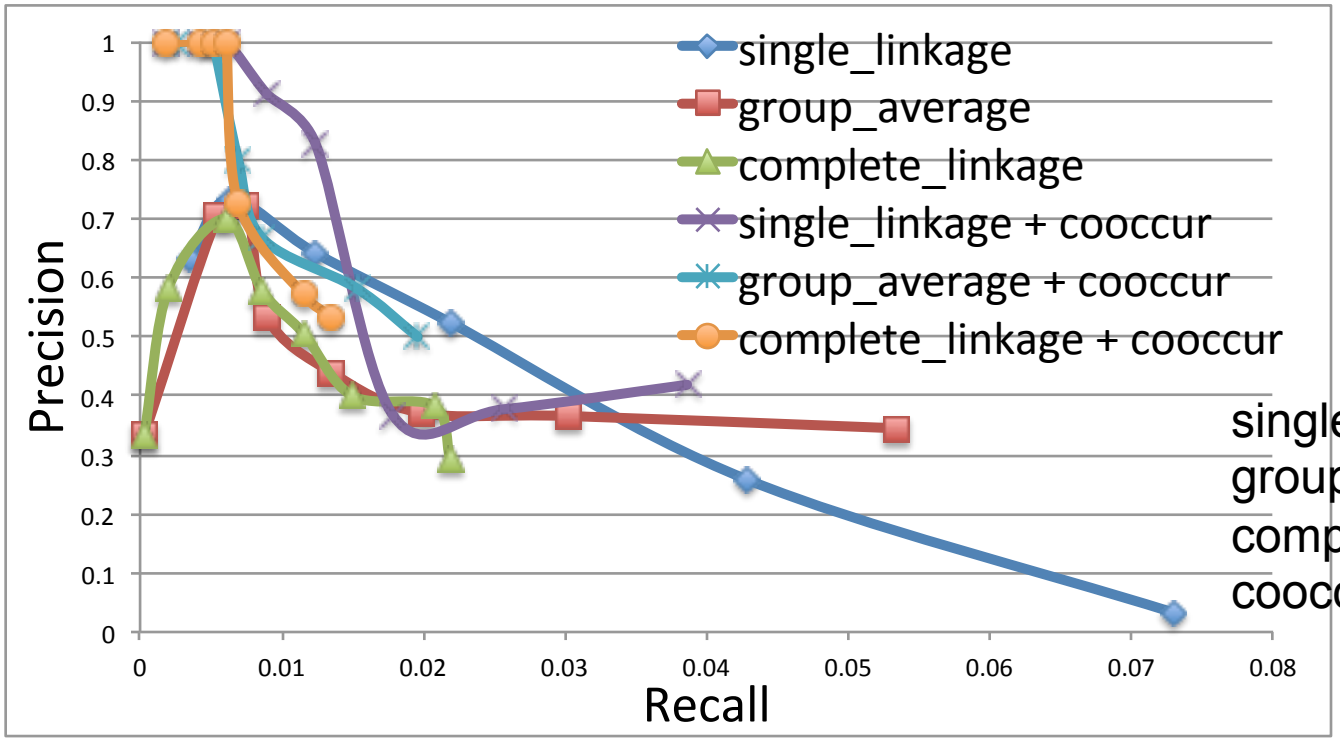
- 階層的クラスタリング
  - 単連結法を用いる
  - p_1とp_2は同義，p_2とp_3は同義 → p_1とp_3は同義
  - co_occur(p_1, p_2)によって同義関係にないパターンは同一クラスタに属さない
  - 要素が2個以下のクラスタは除去
    - 関係を表すクラスタとは考えられないので
- 類似度：共起頻度を素性にしたコサイン類似度

$$sim(p_1, p_2) = \frac{\sum_{j=1} co(p_1, i_j) * co(p_2, i_j)}{\sqrt{\sum_{j=1} co(p_1, i_j)^2} * \sqrt{\sum_{j=1} co(p_2, i_j)^2}}$$

p：パターン
i：インスタンス対
co(p,i)：iとpの共起頻度

実験結果

- コーパス：日本語Webコーパス5.2TB
  - <arg1, arg2, パターン>の3つ組は約32億個抽出される（重複あり）
- カテゴリ対：企業，企業
  - インスタンス数：2705個
  - <インスタンス1, インスタンス2, パターン>の3つ組は672722個（重複あり）
- 正否の判定
  - 全手法での出力から人手で正解データを作成し比較
  - 正解データ：13クラスタ，170個のパターン



single_linkage：単連結法
group_average：完全連結法
complete_linkage：群平均法
cooccur：共起度の制約

獲得した関係の例

single_linkage + pmi, しきい値=0.9でのクラスタリング結果の例

関係	クラスタに含まれるパターンの例	クラスタ数
<企業>が<企業>に売却する	ARG1はARG2に売却することで合意したと発表する ARG1はARG2に売却する方針を固める ARG1はARG2に売却すると発表する	1
<企業>が<企業>を子会社化する	ARG1はARG2を完全だ子会社化する ARG1がARG2を買収する ARG1がARG2を子会社化する	3
<企業>が<企業>の筆頭株主になる	ARG1はARG2の筆頭株主になる ARG1はARG2の筆頭株主となる ARG1はARG2に出資する	1
<企業>が<企業>に負ける	ARG1はARG2に競り負ける ARG1はARG2を撃破 ARG1はARG2と対戦します	2

赤字：誤りパターン

総クラスタ数：28

誤り分析

- 反意の関係にあるパターンが同じクラスタに属してしまう
  - ARG1はARG2に競り負ける，ARG1はARG2を撃破
  - 同じargument対と共起するため
    - <トヨタ自動車，九州電力>，<日本ハム，ソフトバンク>，…
- 因果関係にあるパターンが同じクラスタに属する
  - ARG1はARG2に出資する，ARG1はARG2の筆頭株主となる
  - 文中での共起がないため

これから

- 関係を明示的に表す表現を利用し，関係分類の精度を上げる
  - 接続詞「しかし」や「なので」などとの共起頻度を元に反意の語や因果関係にあるパターンの分類をする
- 得られた関係間の関係（因果関係や反意）の特定を行う