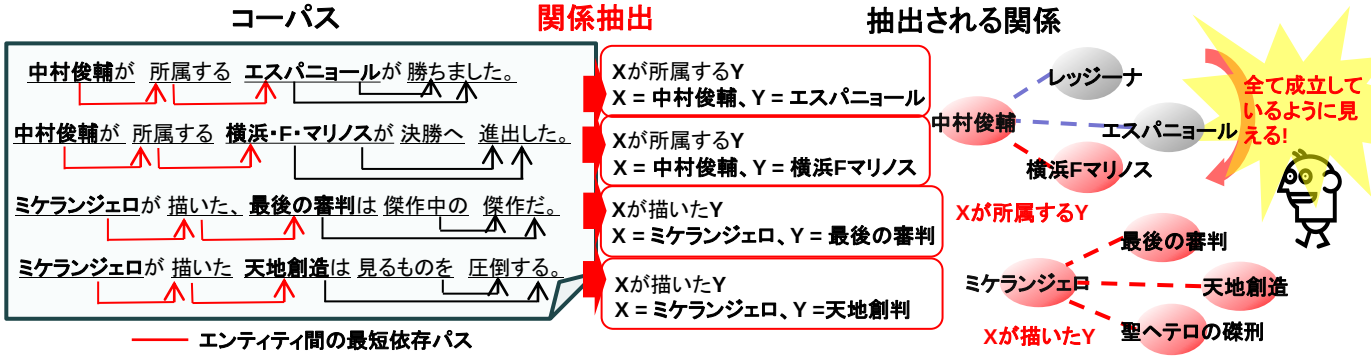
時系列ウェブコーパスを用いた関係抽出の精緻化に関する一考察

東京大学大学院　情報理工学系研究科　高久陽平　鍛治伸裕　吉永直樹　豊田正史

■ 背景

ウェブからの知識獲得として、エンティティ間の関係を抽出する研究が盛んに行われている。しかし、複数のエンティティと同じ関係が抽出されたとき、それが**時系列上の変化によるもの**なのか、もともと**多数のエンティティと関係を持つ**ためなのか判断することができなかった。そこで、我々は**時系列ウェブコーパスを用いることで、関係を精緻化する手法**を考察した。

■ 問題



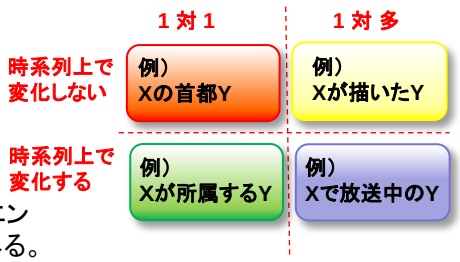
*** 関係の持つ性質を事前に把握しておく必要がある**

■ アプローチ

■ タスク

ウェブから抽出した関係に対し、(一方のエンティティを固定した時)以下の2点から、4値へ分類する。

1. **時系列において変化するか否か**
2. **他方のエンティティは複数存在するか否か**



■ 実験

月別に蓄積したウェブコーパス(ブログ)から抽出した関係に対し、片方のエンティティ(X)を固定したときのもう一方のエンティティ(Y)のとる値の分布を調べる。

■ 実験データ

期間: 2009年1月～2010年12月　　記事数: 8,000万記事 (720 GB)

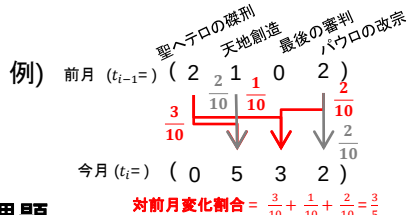
■ 指標

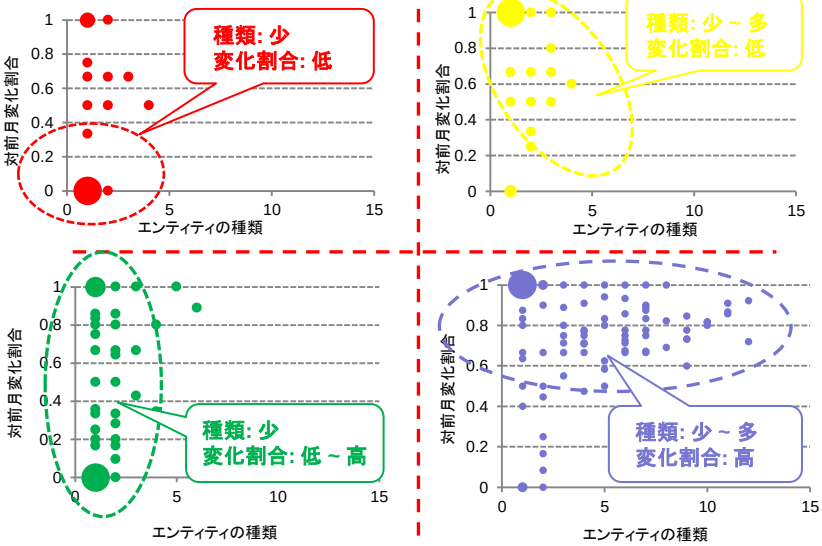
1. **エンティティの種類**
   = Yがとるエンティティの種類数
2. **対前月変化割合**
   = 前月に対して、どれだけエンティティの割合が新たに変化したか

$$= Sum^{+only}\left(\frac{t_i}{Sum(t_i)} - \frac{t_{i-1}}{Sum(t_{i-1})}\right)$$

*t_i はある月iに、Yがとるエンティティの抽出回数をベクトル表現したもの
*Sum^{+only}は与えられたベクトルの正成分だけ合計する関数

例)

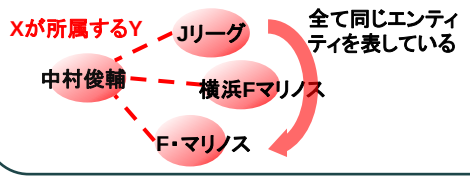




■ 課題

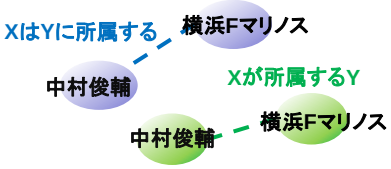
■ エンティティの上位下位概念語、同類語

例) Jリーグ、横浜Fマリノス



■ 語彙統語パターンの表記ゆれ

例) XはYに所属する、Xが所属するY



■ 過去に関する記述の扱い

例) かつて**日本**の首都は**京都**だった