時系列テキストを用いた関係抽出の精緻化

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

・関係抽出は、大規模なWebテキストからの知識獲得の1つとして研究されている

テキストコーパス（→はエンティティ間の最短依存パス）

アップルの → CEOである → スティーブ・ジョブスが → 講演した。

関係抽出

XのCEOであるY
X: アップル、Y: スティーブ・ジョブス

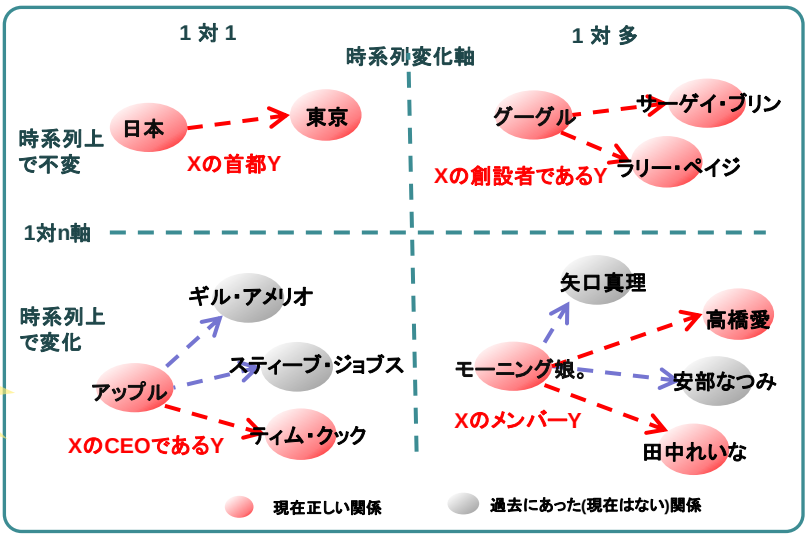
■ 問題点

・ 現在成立している関係と過去に成立していた(現在は成立していない)関係の区別がつかない

→ Q&Aシステムなどにとって重要な問題

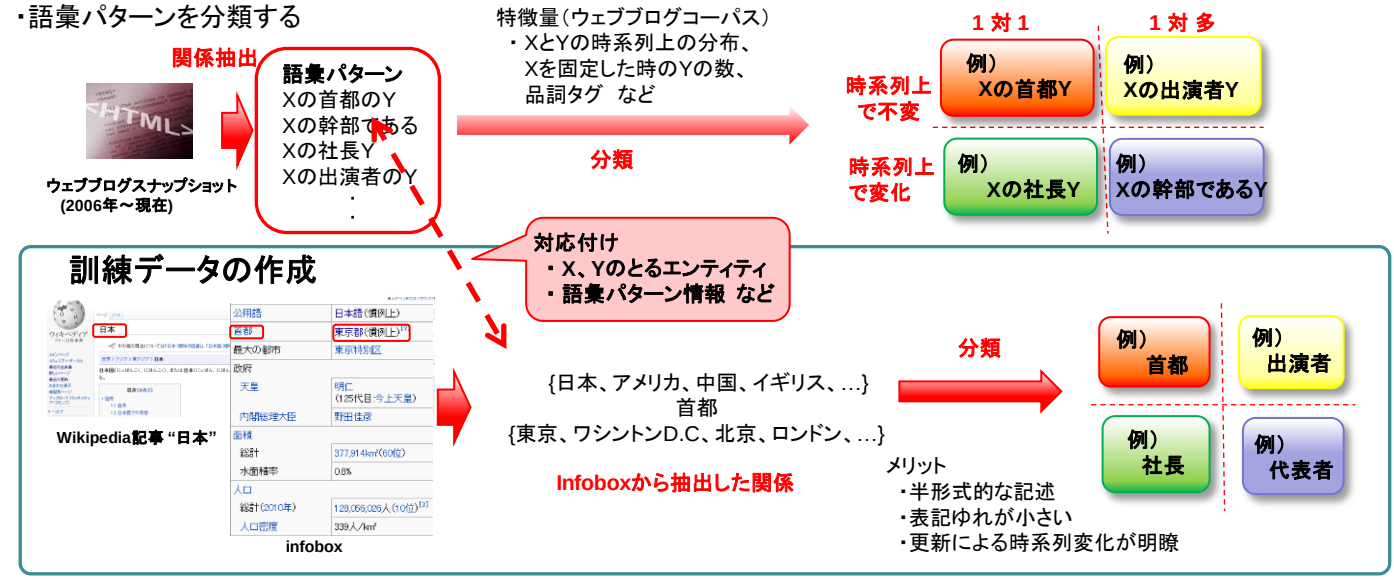
■ 研究目的

・関係を以下の2軸において4つに精緻化
① 時系列変化軸: 時系列上で変化するか否か
② 1対n軸: 関係が1対1か1対多か



■ 分類手法の概要

・語彙パターンを分類する



■ Infobox属性名の分類

① 分類方法

時系列変化軸:Yが編集、更新されるXの割合が
全体のα (=10)% 以上のとき: 時系列変化あり
それ以外: 時系列変化なし

1対n軸:Yを複数とるXの割合が
全体のβ (= 20)% 以上のとき: 1対多
それ以外: 1対1

② 実験データ

・日本語版Wikipedia編集履歴付(2011年8月6日, 約530GB)
・Infobox数: 約 600万個 (約6GB)
・属性名数: 約1 万2,000個
・うち1,000個の属性名を人手でラベル付

③ 実験結果 (精度)

分類軸	ベースライン	提案手法
時系列変化軸	71.9%	**79.1%**
1対n軸	74.3%	**84.9%**
両軸(4分類)	60.6%	**75.0%**

失敗した要因：
・**<松井秀喜、初出場、1993年5月1日>、<松井秀喜、初出場、2003年3月31日>**
→ 現実世界では1対多と捉えたいが、日本、海外でプレーしている選手は稀なため、1対多と認識できない
・**<仁明天皇、子、道康親王>、<仁明天皇、子、時康親王>**
→ "子"が歴史的人物の記事でしか使われないため、時系列変化が認識できない

④ 今後の課題

・infoboxの情報をいかに精度よく形式化できるか
・語彙パターンの対応付けによる分類