[S1-P09] 高精度な流行語予測に向けて：n-gramトレンド分析における同型類語の語彙別成分分解

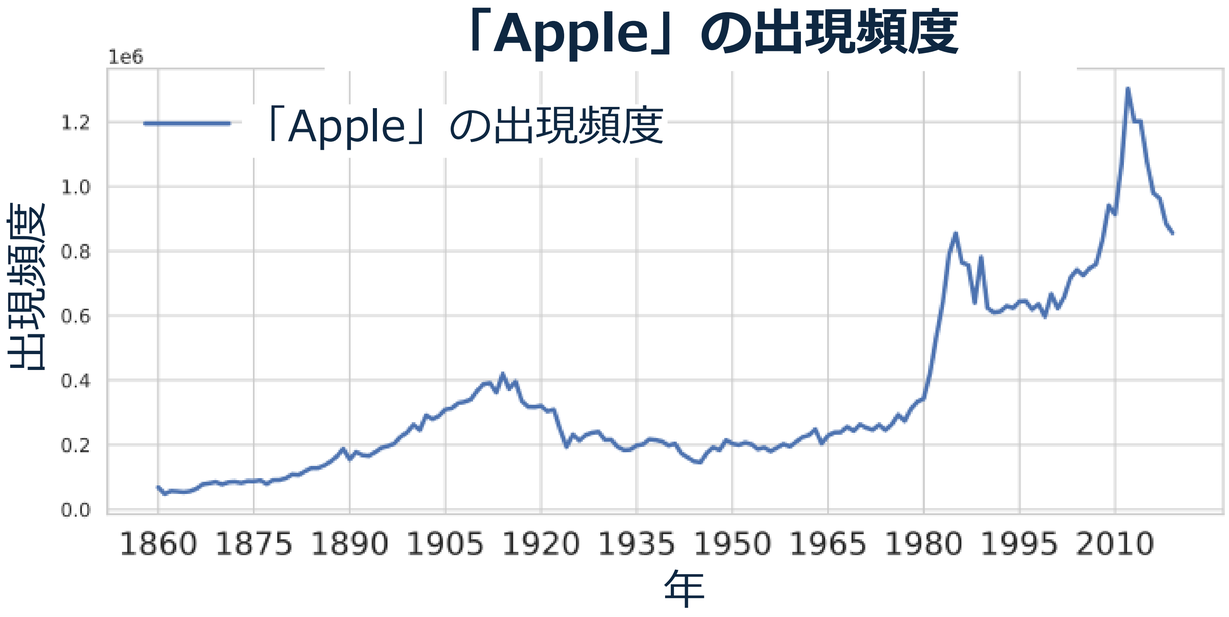
神野 倫行[1]　林 和樹[1]　坂井 優介[1]　上垣外 英剛[1]　渡辺 太郎[1]
(1. 奈良先端科学技術大学院大学)

- 課題: n-gramのトレンド線は**複数の語彙的成分が混合**し、**解釈が難しい**
- 手法: n-gramの出現頻度を**カクテルパーティー問題に置き換え**、ICAを用いた
- 結果: **出現頻度が語彙別に分解**されたことを定性的に確認した

背景：n-gramを活用したトレンド分析における課題

課題
n-gram の出現頻度はトレンド解析に広く用いられるが、表層形依存により**同型類語が混在してしまう**。そのため、語の起源の特定や高精度な流行語の把握が困難となる

目的
本研究はn-gramにおける**出現頻度を語義別成分に分解**することを目的とする。



提案手法：カクテルパーティー問題として捉えるn-gramにおける出現頻度

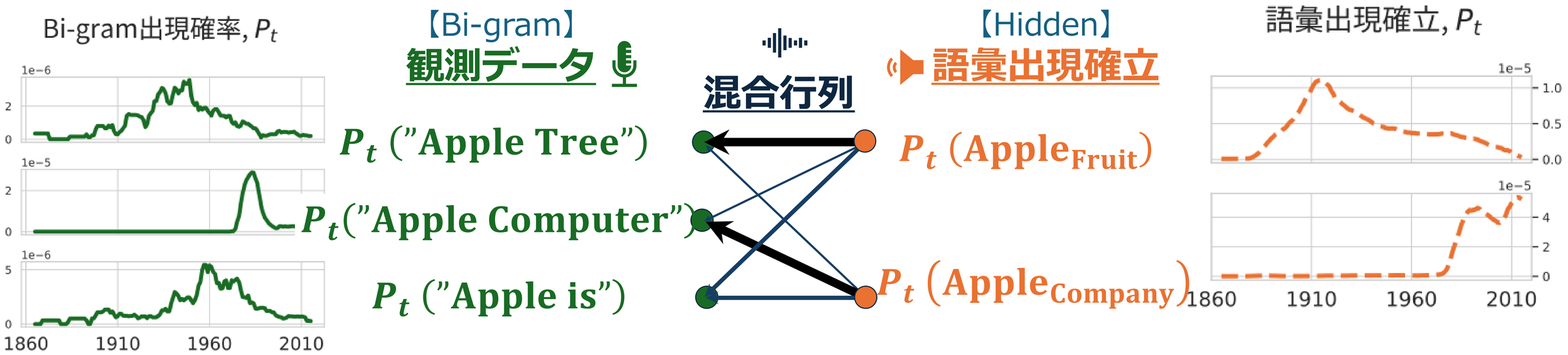
1. 観測データ＝独立成分の線型結合

$$x_t = A\, s_t$$

観測データ【ベクトル】　混合行列　独立成分【ベクトル】

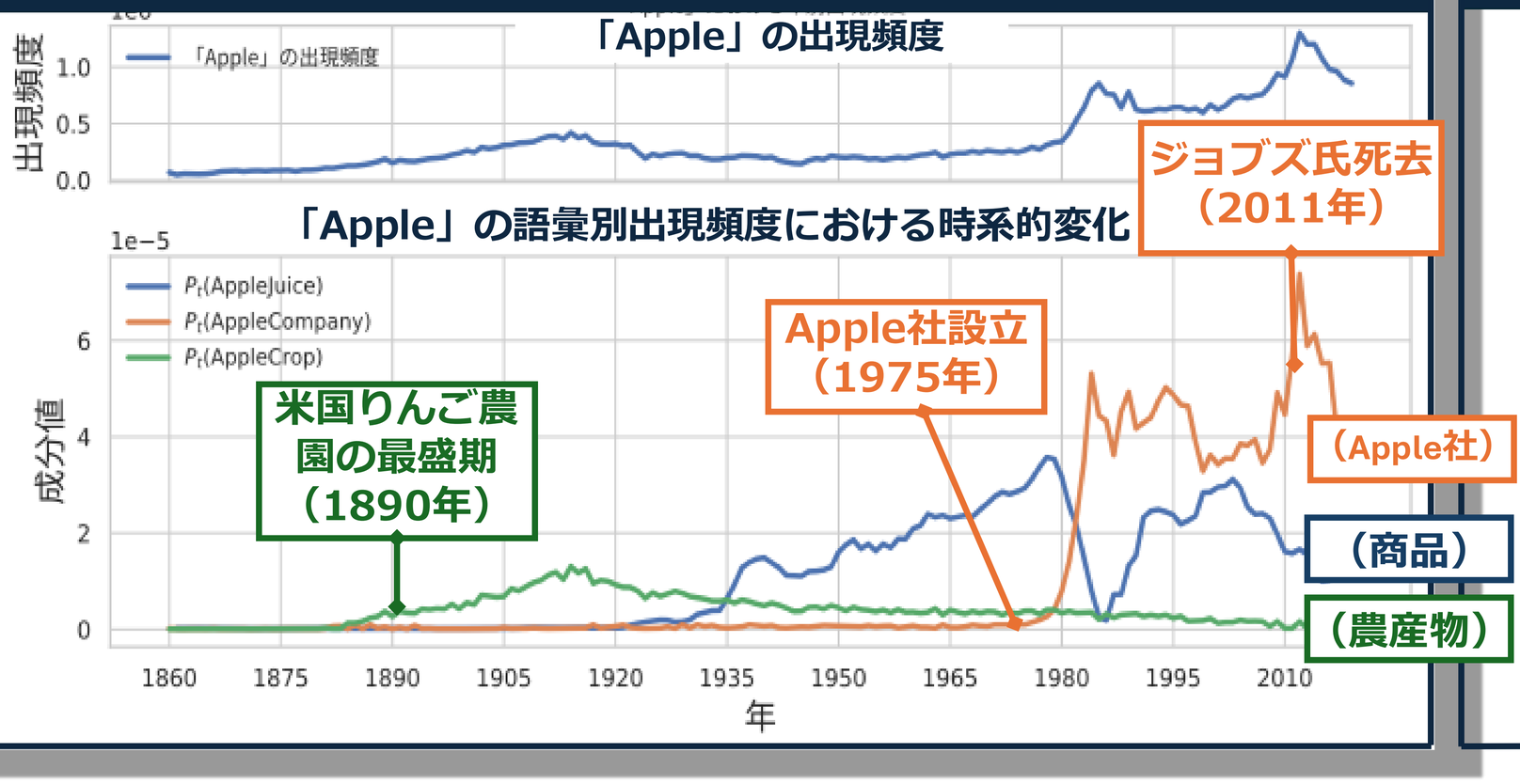
2. n-gramの出現確率を語義成分の線結合と捉える

$$P_t(\text{"Apple Tree"}) = P(\text{"Apple Tree"} \mid \text{Applle}_{\text{Fruit}})\, P_t(\text{Apple}_{\text{Fruit}}) + P(\text{"Apple Tree"} \mid \text{Applle}_{\text{Company}})\, P_t(\text{Apple}_{\text{Company}})$$



独立成分分析（ICA）：Zipff's Law考慮のためベータ事前分布付き二項 MAP ICAを活用。

結果：トレンドデータを語彙別の成分に分解



結論・今後の展望

ICAの活用により、n-gramのトレンド線を**語彙別に分解することが可能であることを**ことを定性的に確認した。

今後はこの様なデータを知識グラフに含めることにより、言葉の**関係性を考慮した、トレンド線の回帰予測**が可能なのかを検証して行く。