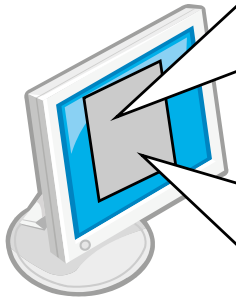
テキスト情報の事実性解析

奈良先端科学技術大学院大学 情報科学研究科
森田 啓　佐尾 ちとせ　松吉 俊　松本 裕治
独立行政法人 情報通信研究機構
乾 健太郎

テキスト情報の事実性

冷やし中華を**食べずにはいられなかった**.



食べずにはいられなかった → 食べた

事実として,
「この著者は冷やし中華を食べた」

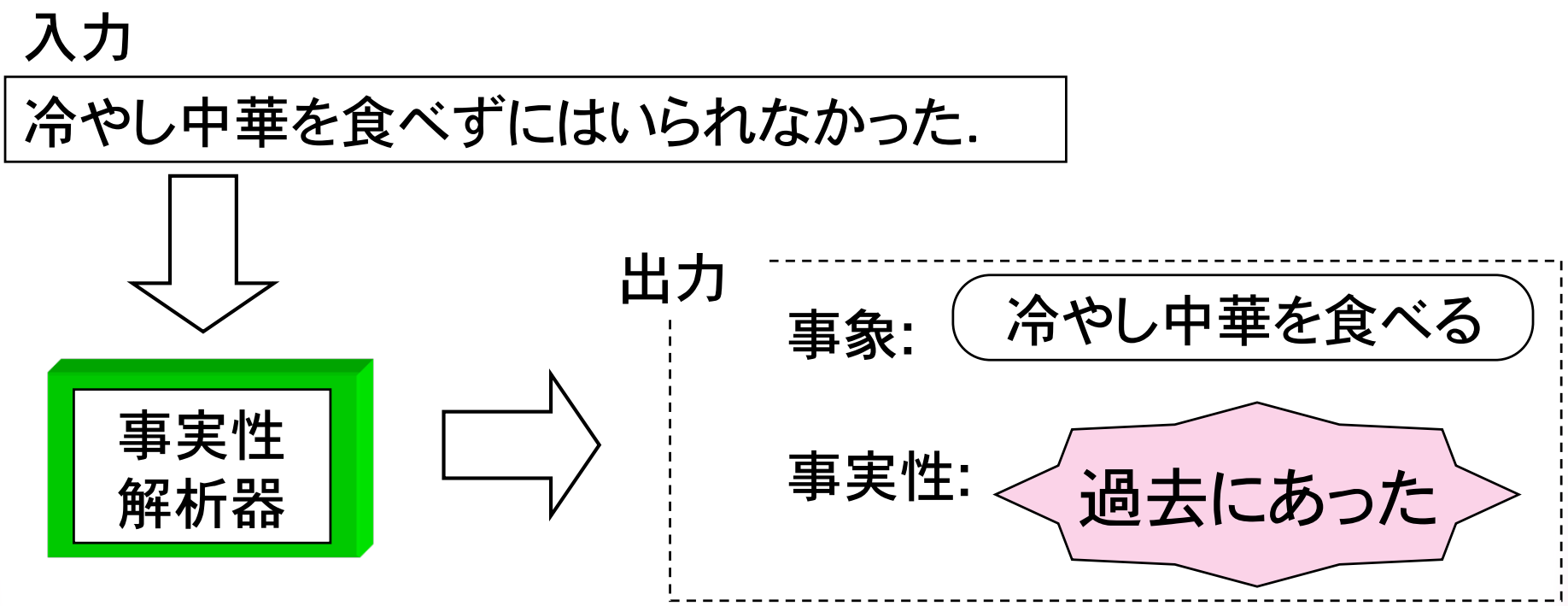
前使ってたシャンプーが余ってしまっている.

前使ってた → 今使っていない

事実として,
「この著者はそのシャンプーを現在使っていない」

本研究の目的

- テキストに含まれる事象の
  事実性を解析するシステムを実現する



事実性解析の難しさ

1. モダリティー表現の存在
   - 推量,仮定などの話者態度が含まれている

お茶のほうが**飲みやすそうですねぇ・・**.

推量

2. 言語表現の多様性
   - 実際のテキストでは,様々な表現が使われる

行くって言ってました

行くつもりだってさ

行こうって話してましたよ

行くらしいと聞いた

行くそうです

行く予定らしいよ

事実性タグ

- 本研究では,事象に対する事実性を
  次のようなタグで表現する

(・,・,瞬間), 意志・予定,(瞬間,・,・)

事象の時間情報
(過去,現在,未来)

モダリティー

モダリティーの時間情報
(過去,現在,未来)

  - 事象の時間情報とモダリティーの時間情報は
    <u>一致するとは限らない</u>

弁当を**買うつもりだったんだけどなぁ**.

事実性タグ体系(原ら 2008を精緻化)

- 時間情報タグ(7種類)
  - 瞬間, 状態, 反復/継続, 始, 止, 否定,
    - (言及なし)
- モダリティータグ(11種類)
  - 事実確定(2種類)
    - 断定, 伝聞
  - 不確定(7種類)
    - 仮定, 推量, 疑い, 質問, 意志・予定, 依頼・当為
  - 無関係(2種類)
    - 体, 用

事実性タグの例

朝,ジュースを**飲んだ**.

(瞬間,・,・),断定,(・,状態,・)

クーラーを**買い替えることにした**.

(・,・,瞬間),意志・予定,(始,状態,・)

東京も雨が**降っているっておばさんが言ってた**.

(・,状態,・),伝聞,(瞬間,・,・)

これから**通うつもりでしたら**,・・・.

(・,・,反復/継続),仮定,(・,状態,・)

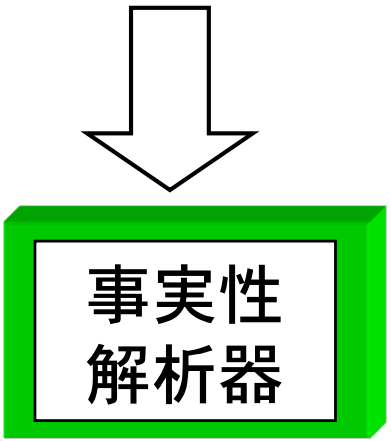
教師あり機械学習を利用

- 事実性に関する言語表現は多様であり，
  これらをすべてあらかじめ列挙しておくことは，
  ほとんど不可能である
- 学習コーパスを作成し，
  事実性タグを機械学習する
  - 2,646文に存在する4,417の事象に対して，
    先に述べた事実性タグを付与した
    - 1人の作業者が作業を行った
    - 別の作業者との一致度(κ統計量)は 0.68

事実性解析の流れ

入力

前使ってたシャンプーが余ってしまっている.



出力

(継続,否定,・),断定,(・,状態,・)

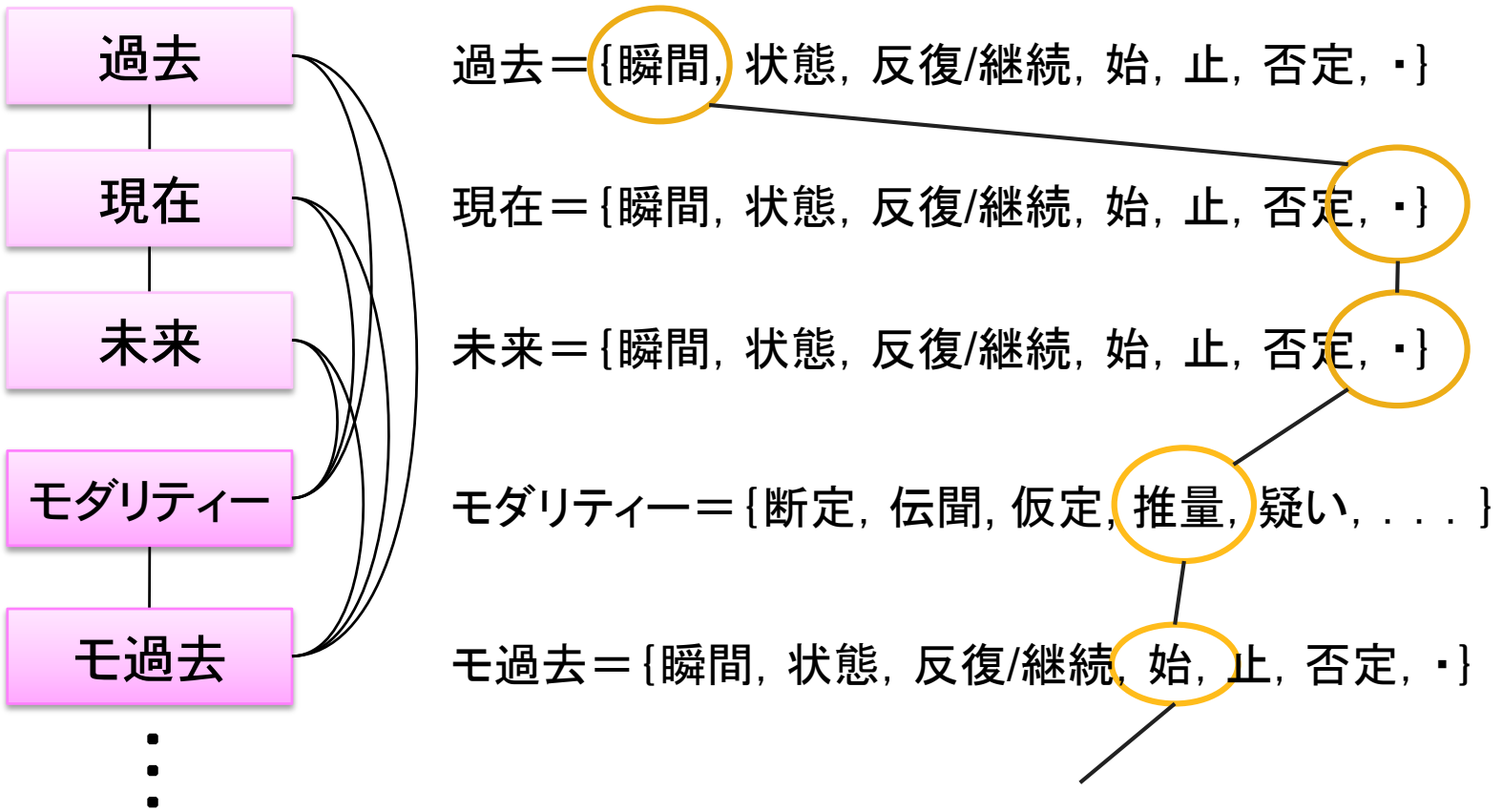
学習

学習コーパス

本研究では,
学習モデルとしてFactorial CRFを利用

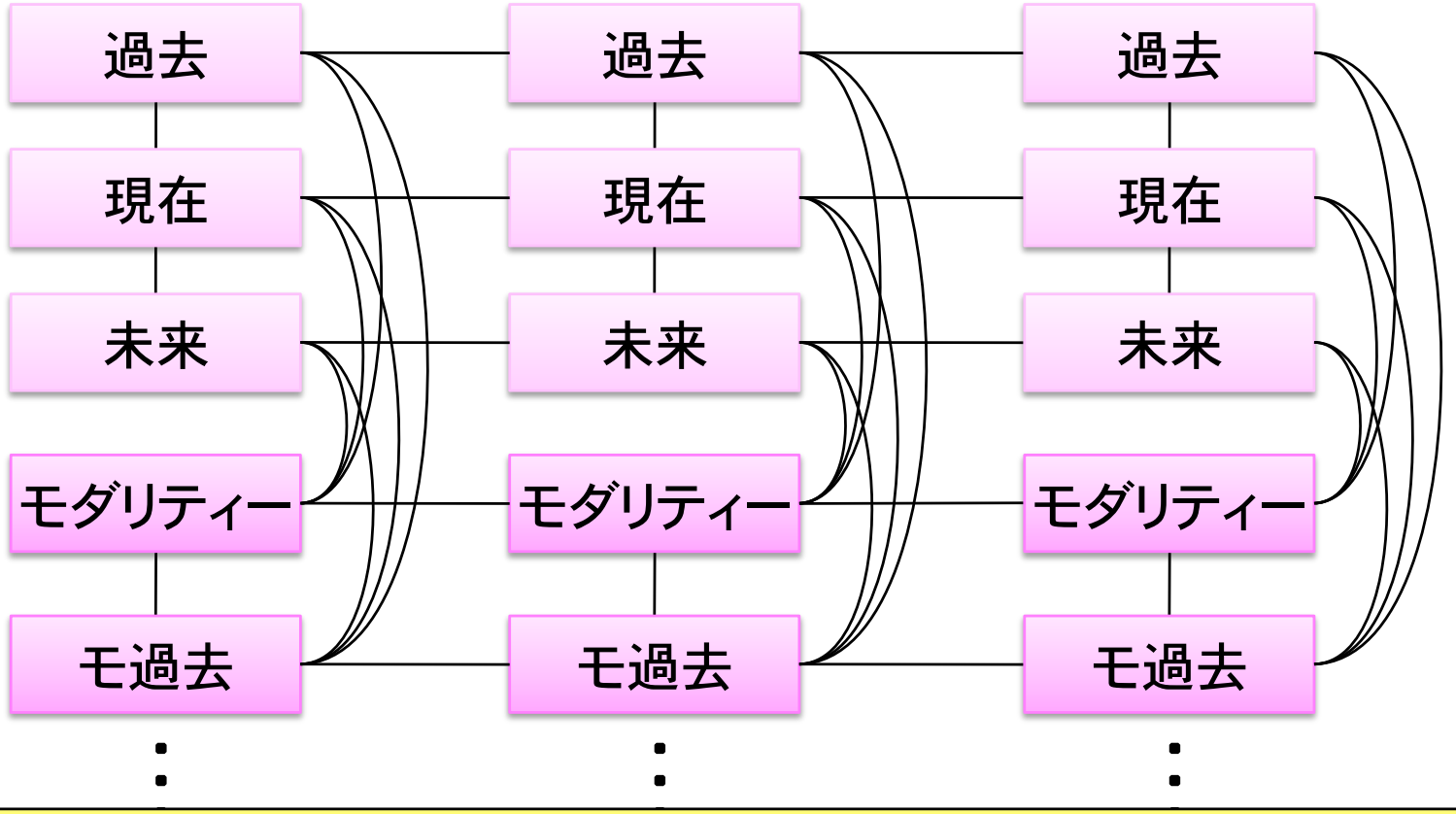
CRF(条件付き確率場)

でもそれを飲みすぎておなかを**壊し**た人の話を聞いてて・・・.



Factorial CRF (Sutton 2006)

でもそれを**飲み**すぎておなかを**壊し**た人の話を**聞い**てて・・・.



同一文内の複数事象間の依存関係を考慮した解析を行う

実験

- 実装した事実性解析器を用いて解析実験を行った
- 使用したFactorial CRFの素性:
  - 以下に対する形態素情報
    - 述語, 前後の文節, 係り先の文節, 係り元の文節
  - 機能表現の意味分類(松吉ら 2007)
    - (例)「とのこと」→伝聞　　「かもしれない」→推量
- 作成したコーパスを用いた(擬似)3分割交差検定
- モダリティーの時間情報に対する評価は省略した
  - その約99% が“ (・,状態,・)”であったため

実験結果

	事象の時間情報			モダリティー
	過去	現在	未来	
ベースライン	0.58	0.54	0.81	0.66
提案手法	0.75	0.69	0.84	0.73

解析成功例:

でも,このお茶高いから買わない(汗)

(・,状態,・),断定

新製品が出たら,買おうともくろんでいる.

(・,・,瞬間),仮定

追加実験

1. 学習コーパスに大量に存在する「断定」タグを大幅に減らして実験を行った
   - 結果: それ以外のタグに対する精度が向上した
   - 結論: 学習コーパスにおけるタグの偏りを無くすことで精度向上が見込めそうである
2. 様々な素性の組み合わせを用いて実験を行った
   - 追加した素性: 単語バイグラム,単語トライグラム
   - 結果: 精度の向上は見られなかった
   - 結論: 事実性解析に有効な全く新しい素性を考案する必要がある

まとめ

- 新たに構築した事実性タグ体系と,事象間の依存関係を考慮する学習モデルFactorial CRFを利用した事実性解析器を提案した
- 作成したコーパスを用いた3分割交差検定において,提案する解析器は,それほど高い精度を達成することはできなかった

今後の課題

- 学習コーパスに大量に存在する「断定」タグ以外のタグに対して学習コーパスの量を増やす
  - 現在は人手で例文を追加している
  - この例文追加作業の自動化を目指している
- 事実性解析に有効な全く新しい素性を考案する