調理映像の理解に向けたレシピの言語処理

森 信介

京都大学学術情報メディアセンター

2012 年 9 月 25 日

Table of Contents

はじめに

レシピテキストの解析

単語分割

固有表現認識

係り受け解析

述語項構造解析

全体の評価

言語処理と映像処理の統合に向けて

おわりに

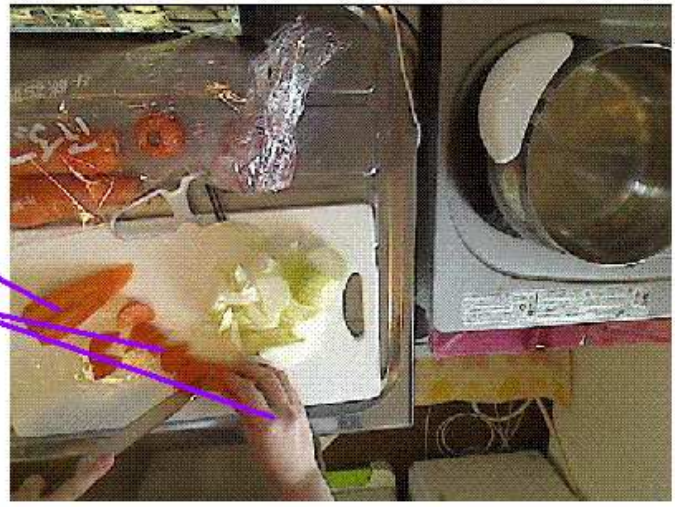
テキストと実施映像

- 指示書 (レシピ) [Momouchi 80]
- 実施映像 (調理風景)

レシピテキスト

1. 玉ねぎ$_F$ 1 個$_Q$ は くし切り$_S$ ， ニンジン$_F$ 1 本$_Q$ と ジャガイモ$_F$ 2 個$_Q$ を 乱切り$_S$ にし$_{Ac}$ ます ． 牛肉$_F$ は 5 cm 程度$_Q$ に 切$_{Ac}$ り ます ．
2. 白滝$_F$ は 下ゆで し$_{Ac}$ て 10 cm 程度$_Q$ に 切$_{Ac}$ って おきます ．

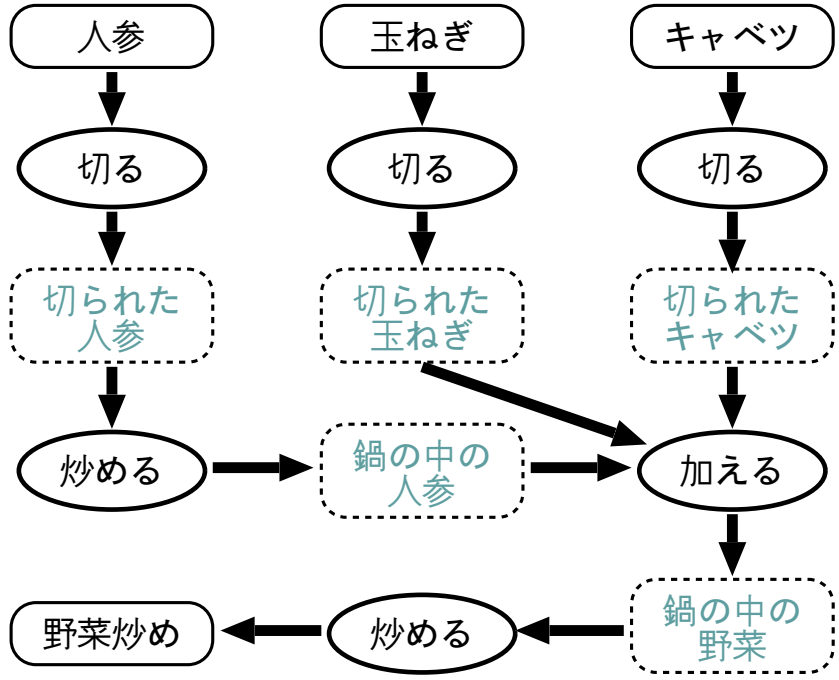
調理映像 (動画)



- 対応を言語・映像の理解と定義

レシピテキスト

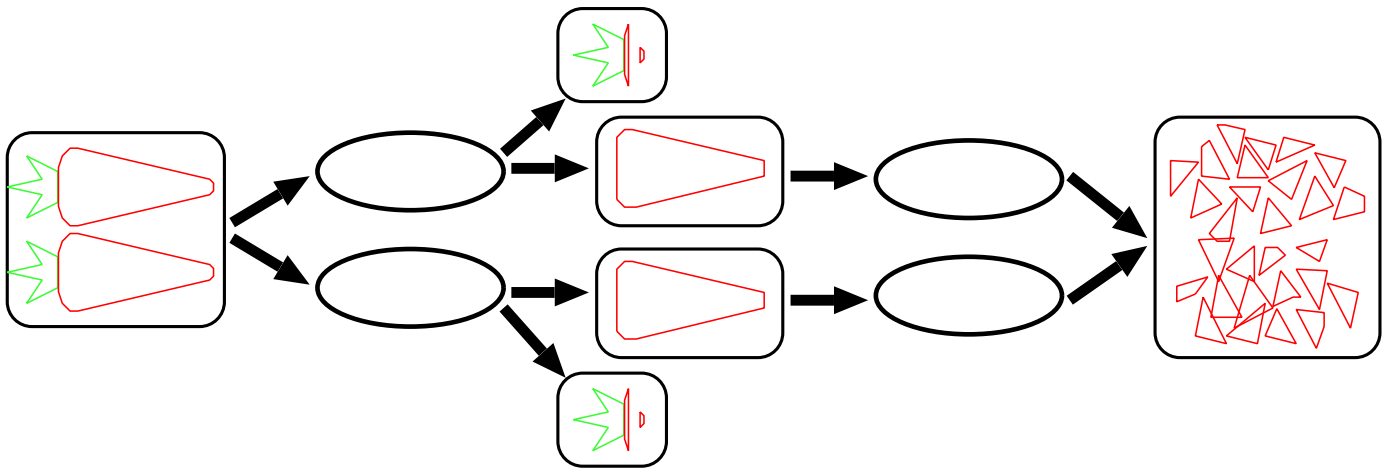
- 文が比較的単純
  - 主観や時制などの問題がほとんどない
  - 言語理解の中間目標
- 有向グラフ (抽象表現) [Hamada 00] [山肩 07]



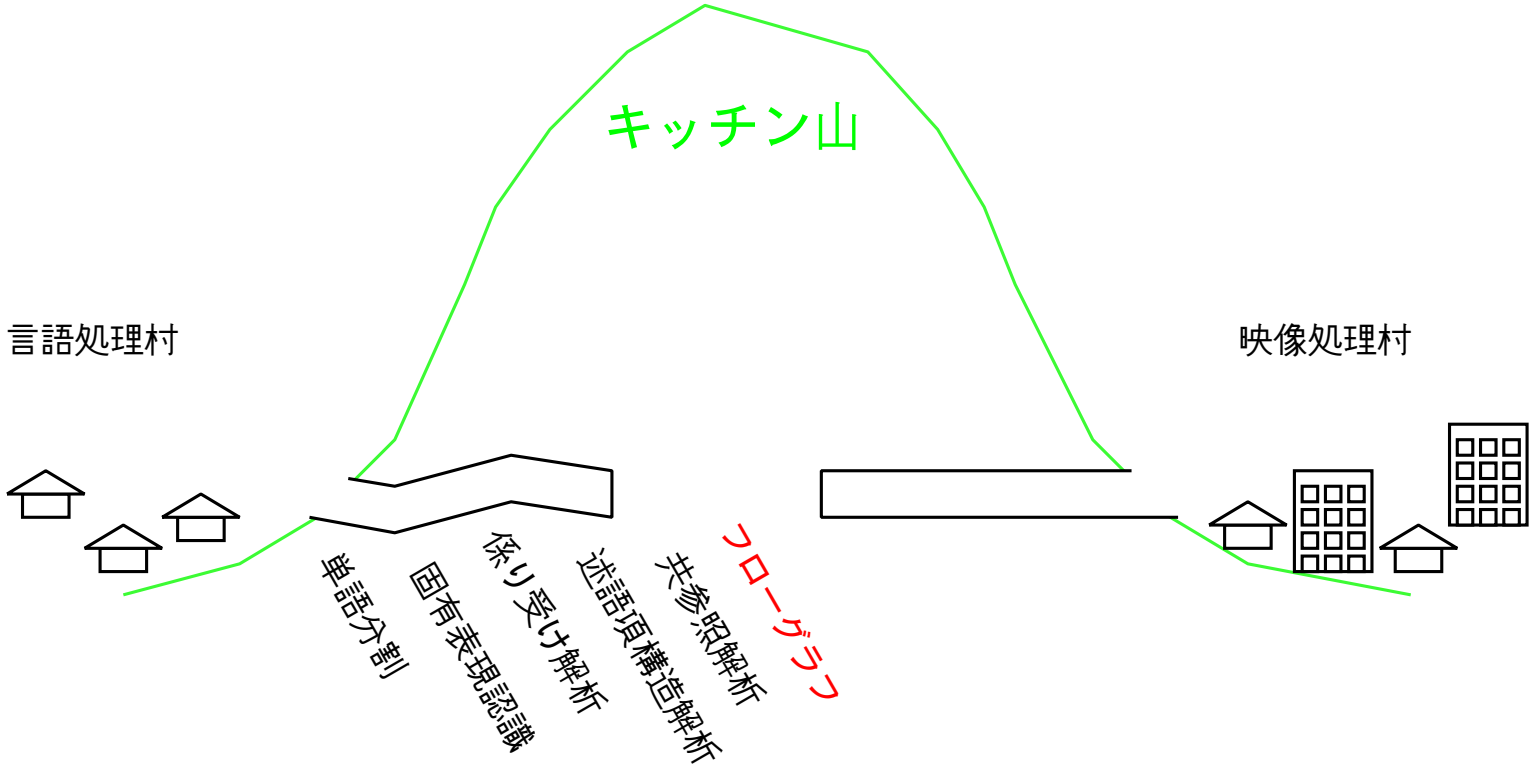
調理映像

- 制御できる環境で進行
  - 光源やカメラの角度が一定
  - 無関係な物体や動きが少ない
- 物体追跡からの有向グラフ
  - アノテーション基準の確立 [橋本, 舩冨, et al.]

例 人参を乱切りにする



レシピと調理映像のマッチング



- 一般的工法
- 紆余曲折

テキスト解析

最先端の言語処理 + 分野適応

1. 単語分割 [Neubig, Mori, et al. 11]
   - 文中の単語の認定
   - 活用語の原形推定を含む
   - KyTea (きゅーてぃー) (Cf. 茶筌, MeCab, JUMAN, ...)
2. 固有表現認識
   - 実世界の物体や行動に対応する単語列
   - 種類は独自設定
     食材 (F), 量 (Q), 道具 (T), 継続時間 (D),
     食材の状態 (S), 調理者の動作 (Ac), 食材の動作 (Af)

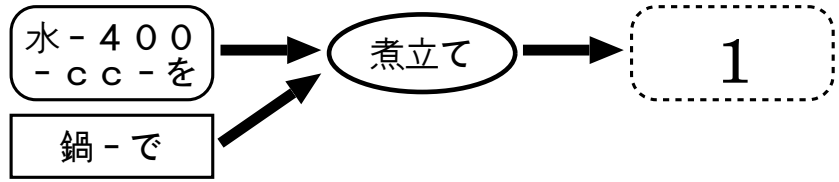
テキスト解析 (つづき)

3. 係り受け解析　　[Flannery, Mori, et al.]
   - 単語や固有表現間の統語的関係
   - EDA (Cf. CaboCha, KNP, ...)

4. 述語項構造解析
   - 単語や固有表現の動作に対する意味的役割
   - 規則に基づく方法 ⇒ 機械学習

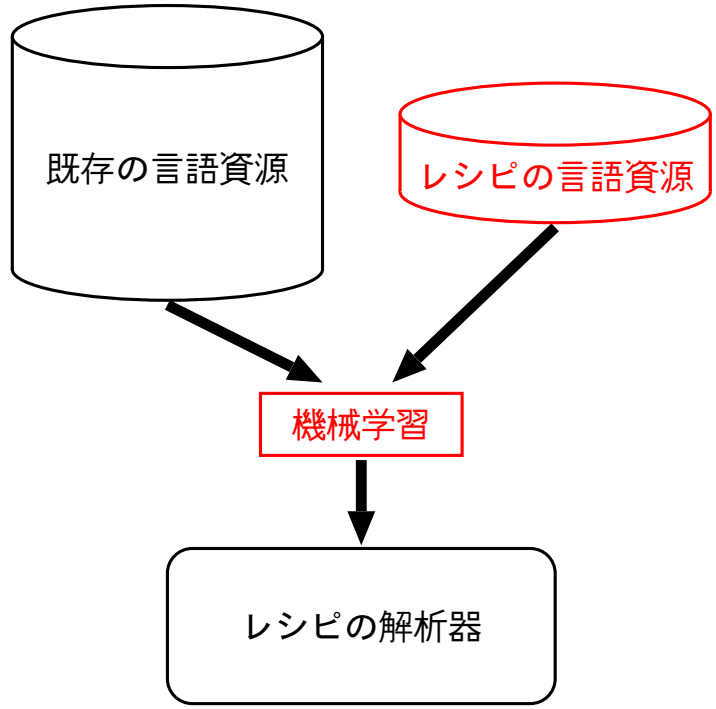
出力

煮立て$_{\mathbf{Ac}}$(を:水-４００-ｃｃ$_{\mathbf{F}}$, で:鍋$_{\mathbf{T}}$)



レシピテキストへの適応

- レシピテキストは単なる一例
  - 一般的な分野適応の方法を追求 [森 12]



- 機械学習部分と適応対象の言語資源を設計

言語資源

- 既存: 一般分野のフルアノテーション

出典	文 数	文 字 数	固有表現数	係り受け数
BCCWJ	53,899	1,834,784	–	–
辞書の例文	11,700	197,941	–	136,109
新聞記事	9,023	398,569	–	254,402

BCCWJ: 現代日本語書き言葉均衡コーパス [前川 09]

- レシピテキスト : フルアノテーション

出典	文 数	文 字 数	固有表現数	係り受け数
固有表現 認識の学習	242	7,023	1,523	–
テスト	724	19,966	3,797	12,426

Step1. 単語分割 (単語の同定)

- 入力: 文
  水４００ｃｃを鍋で煮立て、沸騰したら中華スープの素を加えてよく溶かす 。
- 出力: 単語列

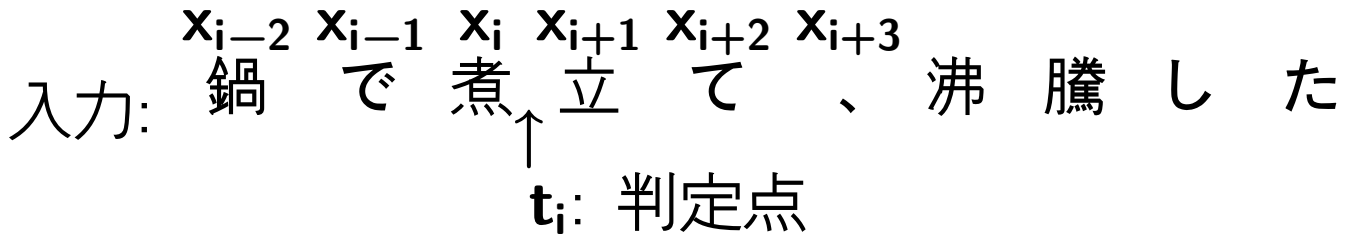
  水|４-０-０|ｃ-ｃ|を|鍋|で|煮-立-て|、|
  沸-騰|し|た-ら|中-華|ス-ー-プ|の|素|を|
  加-え|て|よ-く|溶-か|す|。

  - |: 単語境界あり
  - -: 単語境界なし

※ 活用語尾の分割 ⇒ 活用語の正規化

Step1. 単語分割 (KyTea, 点予測) [Neubig 11]

- 文字間の 2 値分類問題 ⇒ 線形 SVM による実装



- 素性

  Char (type) 1-gram feature:

  -3/鍋 (K), -2/で (H), -1/煮 (K), 1/立 (K), 2/て (H), 3/、(S)

  Char (type) 2-gram feature:

  -3/鍋で (KH), -2/で煮 (HK), -1/煮立 (KK), 1/立て (KH), 2/て、(HS)

  Char (type) 3-gram feature:

  -3/鍋で煮 (KHK), -2/で煮立 (HKK), -1/煮立て (KKH), 1/立て、(KHS)

部分的アノテーションコーパスから学習可能
⇒ 柔軟なコーパス作成!
⇒ 迅速・安価な分野適応!

部分的アノテーションコーパス

- 文は複数の判定箇所を含む
- 一部の判定箇所のみにラベル付与

1. 未知語候補の抽出 [Mori 96]
2. 単語境界の修正作業

玉ねぎ (頻度=1362)

… | 玉 – ね – ぎ | は 薄 切 り 、 ピ ー マ ン は 薄 い 輪 …
… マ リ ネ 液 を 作 り 、 （ 1 ） の | 玉 – ね – ぎ | ・ …
… 約 6 分 加 熱 す る 。 | 玉 – ね – ぎ | は 粗 み じ ん …

こん (頻度=1338)

… 移 し 、 「 | こ – ん – ぶ | だ し 」 、 半 ず り 白 ご …
… 入 れ 、 両 面 を | こ – ん – が – り – と | 色 づ く ま で …
… 2 つ 切 り 、 | れ – ん – こ – ん | は 皮 を む い て 8 …

文脈情報の重要性

- 一般分野から Web(Yahoo!知恵袋) への分野適応
  `http://www.phontron.com/kytea/dictionary-addition.html`
  (2011 年 11 月 25 日)

- 単語分割の精度

モデル	精度 (F 値)
適応なし	95.54%
辞書追加 (文脈なし)	96.75%
コーパス追加 (文脈あり)	97.15%

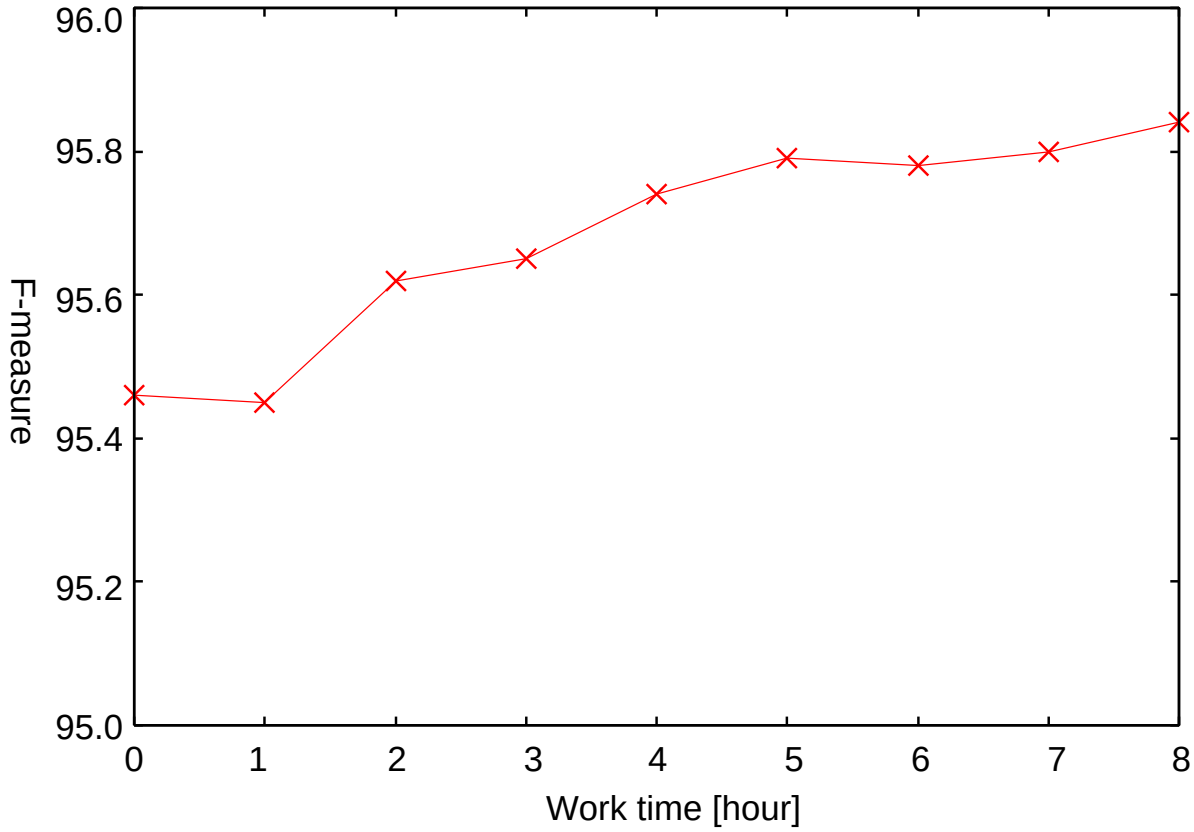
- 約 75〜80%の精度向上は辞書追加により実現可能
  - 多くの言語処理応用ではここまで

- 残りの 20〜25%の精度向上には文脈情報が必要

一般モデルとその分野適応

- 一般モデル: BCCWJ, UniDic, など
- 適応モデル: 未知語候補への部分的アノテーション
  - 8 時間

学習曲線



- F 値: 再現率と適合率の調和平均

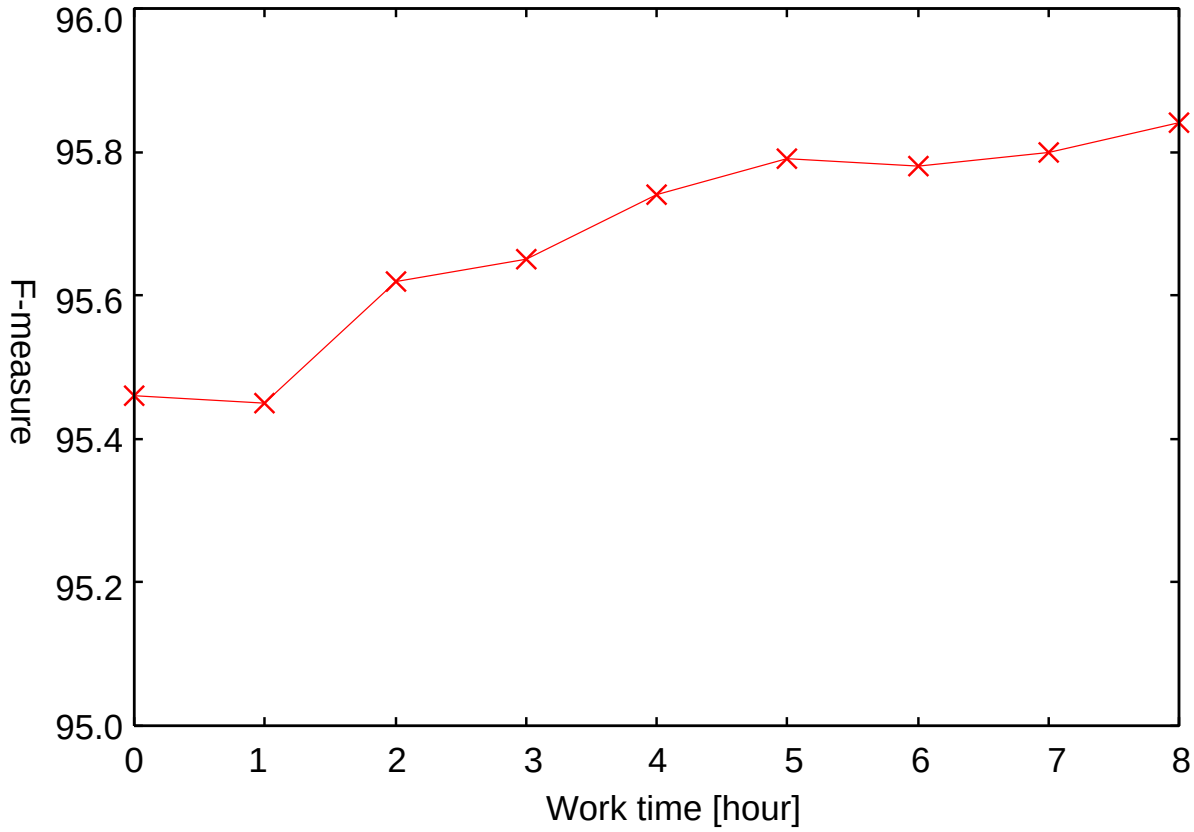
  再現率 ＝ **LCS**/出力

  適合率 ＝ **LCS**/正解

longest common subsequence

※ **LCS**：最長共通部分系列

学習曲線



- 一般モデルでは不十分 (一般分野: 99%程度)
- 作業時間にしたがって精度向上
- さらなる作業が必要

Step 2. 固有表現認識

- 固有表現
  - 実世界の物体や動作に対応する単語列
  - 一般的には、人名、組織名、時間、 ...
  - 定義はタスク依存 $\Rightarrow$ 一般分野コーパスがない
- レシピの固有表現を独自に設定:
  食材 (**F**), 量 (**Q**), 道具 (**T**), 継続時間 (**D**),
  食材の状態 (**S**), 調理者の動作 (**Ac**), 食材の動作 (**Af**)

  水$_{\mathbf{F}}$ ４００ ｃｃ$_{\mathbf{Q}}$ を 鍋$_{\mathbf{T}}$ で 煮立て$_{\mathbf{Ac}}$ 、沸騰し$_{\mathbf{Af}}$ たら
  中華 スープ の 素$_{\mathbf{F}}$ を 加え$_{\mathbf{Ac}}$ て よく 溶か$_{\mathbf{Ac}}$ す 。

点予測による固有表現認識

部分的アノテーションコーパスから学習可能
⇒ 柔軟なコーパス作成!
⇒ 迅速・安価な分野適応!

1. BIO2 表現 (1 単語に 1 つの固有表現タグ)
   水/B-F ４００/B-Q ｃｃ/I-Q を/O 鍋/BT で/O
   煮立て/B-Ac 、/O 沸騰/B-Af し/I-Af たら/O
   中華/B-F スープ/I-F の/I-F 素/I-F を/O 加え/B-Ac
   て/O よく/O 溶か/B-Ac す/O 。/O

2. 部分的アノテーションコーパスから単語のタグを推定するロジスティック回帰を構築
   - 現状では部分的アノテーションコーパスはない
   - Cf. CRF の学習にはフルアノテーションが必要

点予測による固有表現認識 (つづき)

3. 各単語に対して可能なタグと確率を出力

		w			
P(y|w)	水	４００	ｃｃ	を	・・・
F-B	0.62	0.00	0.00	0.00	・・・
F-I	0.37	0.00	0.00	0.00	・・・
Q-B	0.00	0.82	0.01	0.00	・・・
y Q-I	0.00	0.17	0.99	0.00	・・・
T-B	0.00	0.00	0.00	0.00	・・・
⋮	⋮	⋮	⋮	⋮	⋱
O	0.01	0.01	0.00	1.00	・・・

4. 解釈可能な最適タグ列を探索

   例 “**F-I Q-I**” は解釈不可能

初期モデルと分野適応

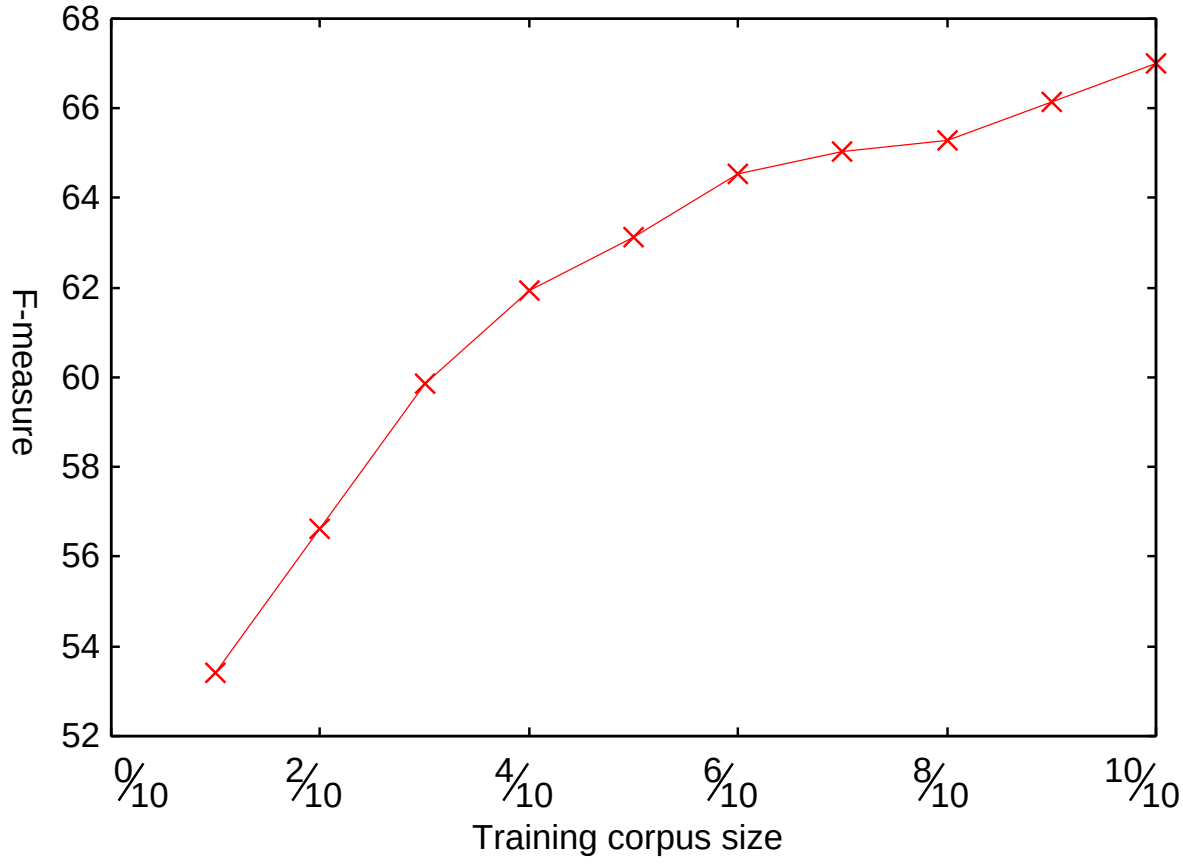
- 肉じゃがのレシピ (242 文) にタグ付与 (5 時間)

  ↑ 良くない設定 ⇒ 無作為抽出に変更中

- 初期モデル: 1/10 を利用
- 適応モデル: 2/10 から 10/10 を利用

学習曲線

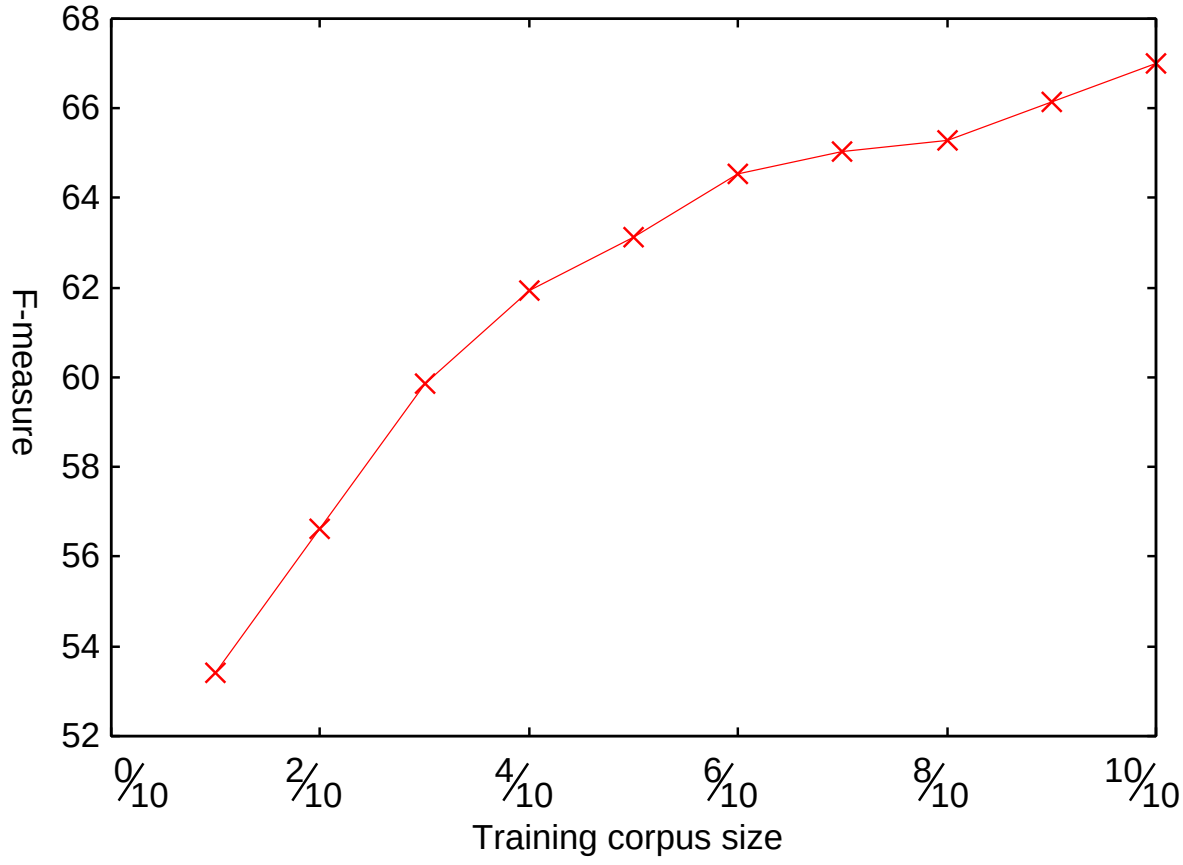
- F 値



- 一般的な固有表現認識タスク (80〜90%) より低い
  - ex. 学習 = 11,000 文で 83.1%, 1,038,986 語で 90.0%)

学習曲線

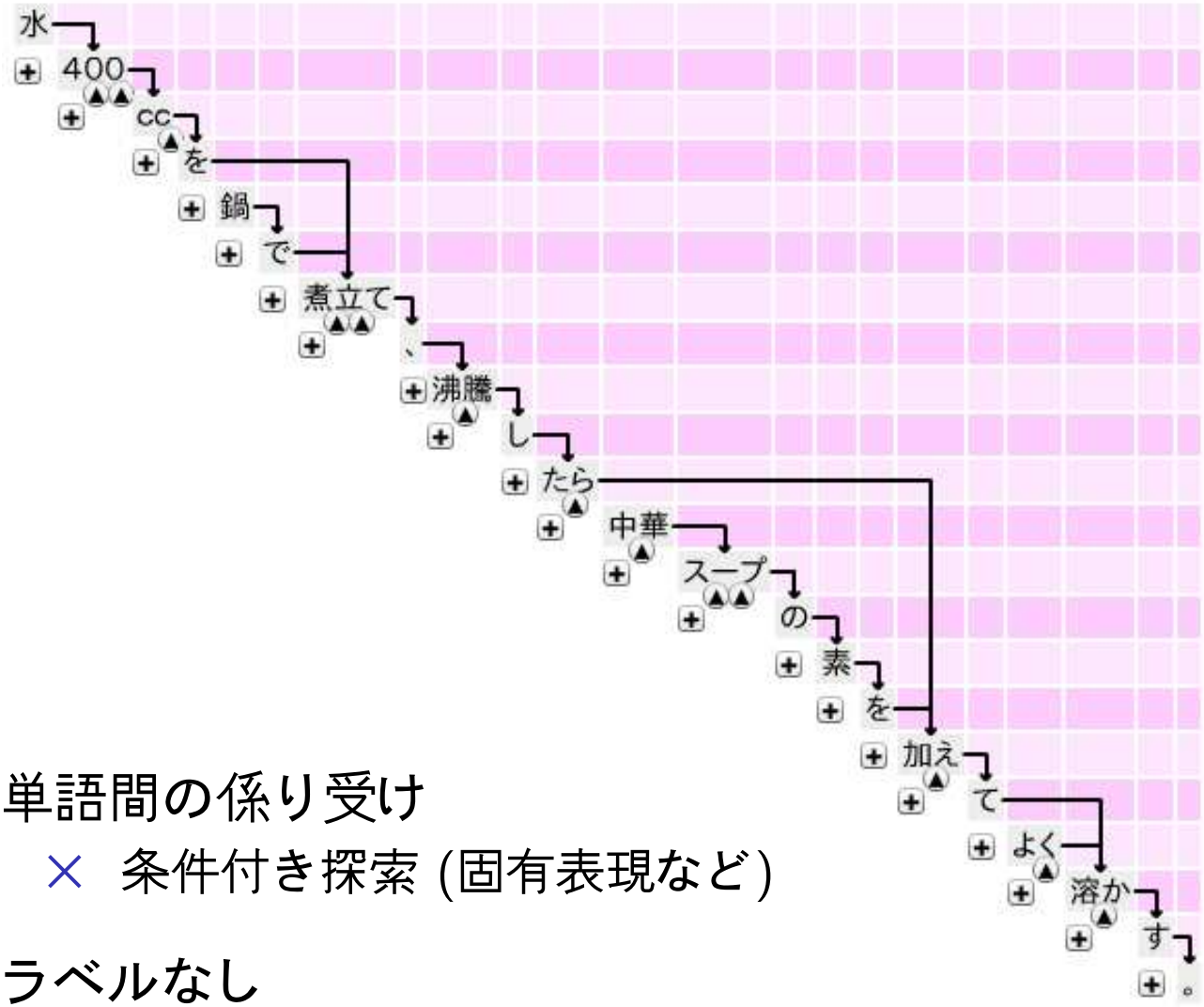
- F 値



- アノテーション作業にしたがって急激に上昇
  - ex. 5 時間 (243 文) ⇒ 250 時間 (12,150 文)

Step 3. 係り受け解析 (EDA, 点予測) [Flannery 11]

- 文の統語構造



  - 単語間の係り受け
    - × 条件付き探索 (固有表現など)
  - ラベルなし

点予測による係り受け解析

- 点予測による最大全域木 (EDA) [Flannery 11]
  1. 全ての単語間の係り受けスコアを計算

$$\sigma(\langle \mathbf{i}, \mathbf{d_i} \rangle, \vec{\mathbf{w}}), \qquad \text{ここで } \mathbf{w_i} \text{ は } \mathbf{w_{d_i}} \text{ に係る}$$

  2. エッジスコアの合計が最大になる全域木 (MST) を選択

$$\hat{\vec{\mathbf{d}}} = \underset{\vec{\mathbf{d}} \in D}{\mathbf{argmax}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \sigma(\langle \mathbf{i}, \mathbf{d_i} \rangle, \vec{\mathbf{w}})$$

部分的アノテーションコーパスから学習可能
⇒ 柔軟なコーパス作成!
⇒ 迅速・安価な分野適応!

点予測による係り受け解析 (つづき)

- スコア計算の素性

牡蠣 を 広島 に 食べ に 行 く

w_{i-3} w_{i-2} w_{i-1} w_i w_{i+1} w_{i+2} w_{i+3}

w_{d_i-3} w_{d_i-2} w_{d_i-1} w_{d_i} w_{d_i+1} w_{d_i+2} w_{d_i+3}

F1 係り元 w_i と係り先 w_{d_i} の距離
F2 w_i と w_{d_i} の表記
F3 w_i と w_{d_i} の品詞
F4 w_i と w_{d_i} の前後 3 単語の表記
F5 w_i と w_{d_i} の前後 3 単語の品詞

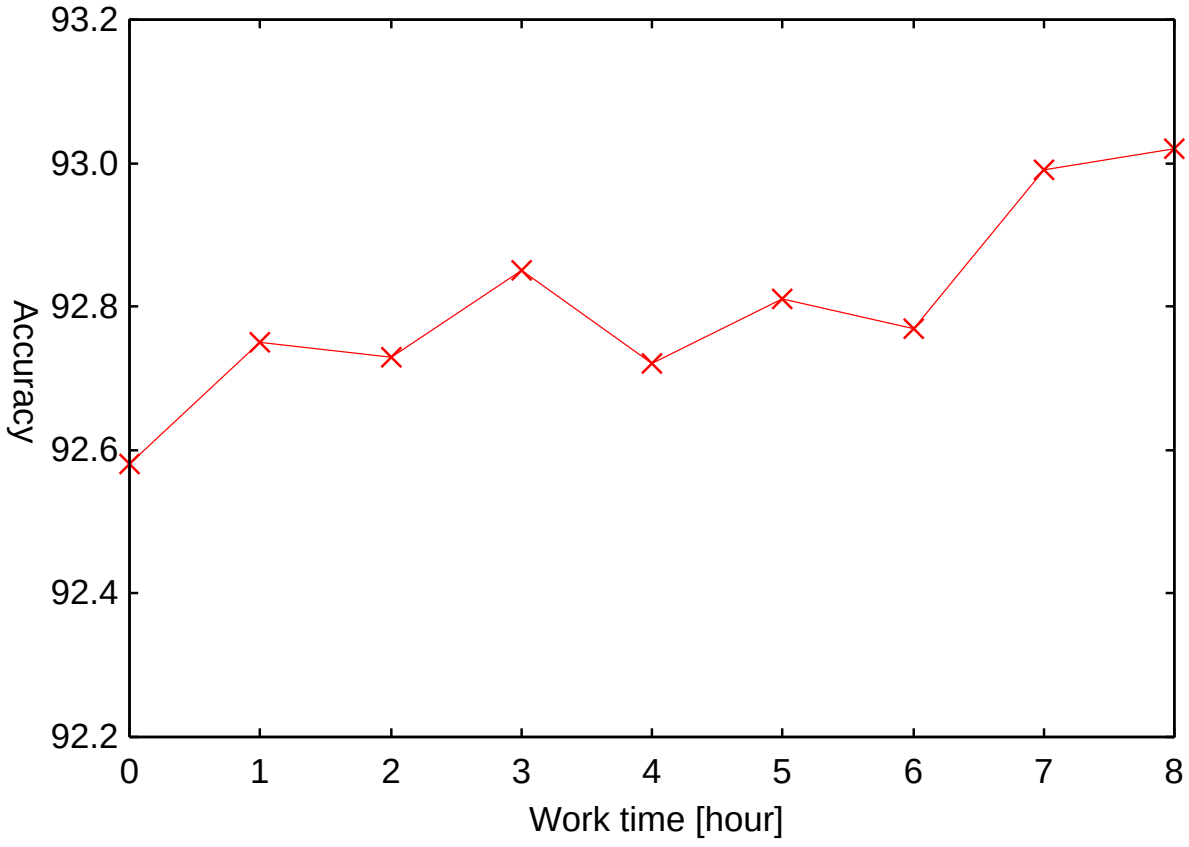
一般モデルとその分野適応

- 一般モデル: 約 2 万文から学習
  - EHJ (英語表現辞典の例文): 11,700 文, 145,925 語
  - NKN (日経新聞の記事): 9,023 文, 263,425 語
- 分野適応: 新出の名詞と助詞の組に係り先を付与
  1. 既存のアノテーションに含まれない名詞と助詞の列を見つける
  2. 名詞から用言までの係り受けを付与する

     ｃｃ → を → ... 煮立て
  3. 8 時間の作業

結果

- 学習曲線



- 一般分野に対する精度 (96.83%) と比べて低い
- 作業時間にしたがって精度向上

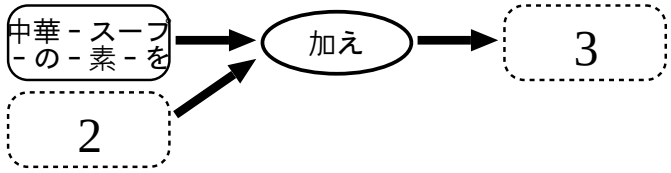
Step 4. 述語項構造解析

- 現状は規則に基づく方法
  - 点予測による機械学習に変更
  - ゼロ照応などの復元
- 有向グラフの最小の単位に対応

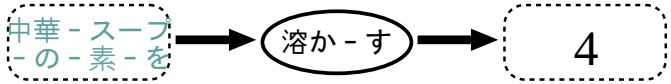
1. 煮立て$_{\mathbf{Ac}}$(Chef, 水$_{\mathbf{F}}$ ４００ ｃｃ$_{\mathbf{Q}}$ を, 鍋$_{\mathbf{T}}$ で)

2. 沸騰-し$_{\mathbf{Af}}$(Food), たら

3. 加え$_{\mathbf{Ac}}$(Chef, 中華 スープ の 素$_{\mathbf{F}}$ を, 水$_{\mathbf{F}}$ に)



4. 溶か-す$_{\mathbf{Ac}}$(Chef, 中華 スープ の 素$_{\mathbf{F}}$ を)



機械学習による述語項構造抽出

- 動的素性を使わない設計

部分的アノテーションコーパスから学習可能
⇒ 柔軟なコーパス作成!
⇒ 迅速・安価な分野適応!

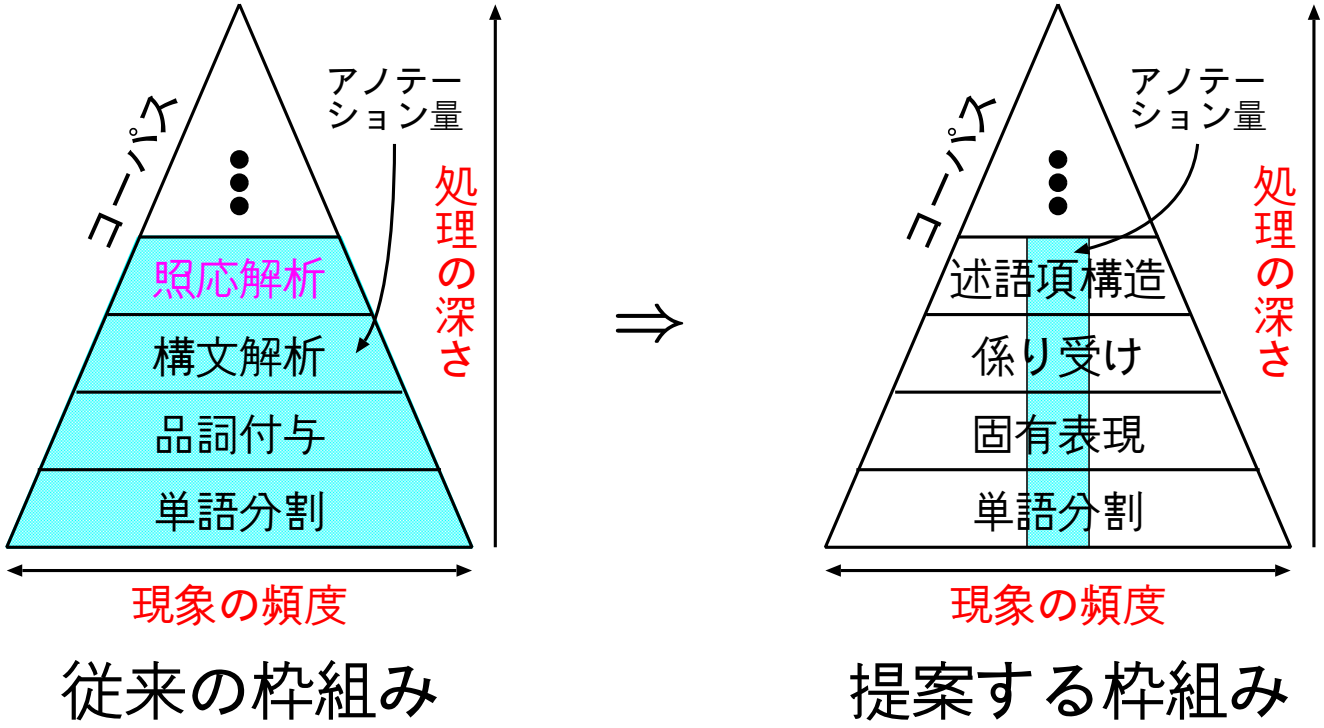
全体の評価

1. テスト コーパス: 無作為抽出の 100 レシピ

出典	文 数	文 字 数	固有表現数	係り受け数
テスト	724	19,966	3,797	12,426

2. 学習コーパス
   - 単語分割:
     (BCCWJ + etc.) + 部分的アノテーション
   - 固有表現認識:
     肉じゃが 1/10 + 9/10 (設定が良くない)
   - 係り受け解析:
     (辞書の例文 + 新聞記事) + 部分的アノテーション
   - 述語項構造解析: 規則による方法 ⇒ 機械学習

各段階の言語資源を独立となるように設計



従来の枠組み　　提案する枠組み

- 点予測で容易に実現
- (統一的の) 系列予測でも実現可能のはず
  - 異なる処理段階の統一は昔から課題

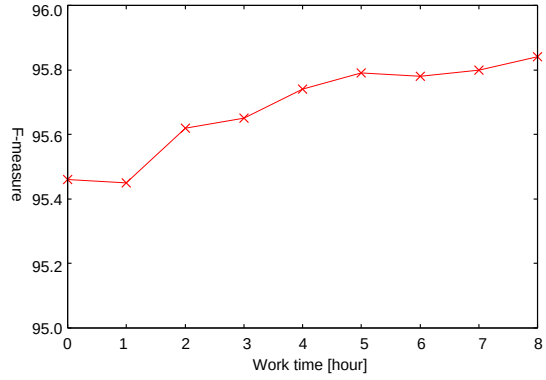
各処理の結果のまとめ

Step 1. 単語分割

一般モデル: 95.46%

⇓ (8 時間)

分野適応後: 95.84%

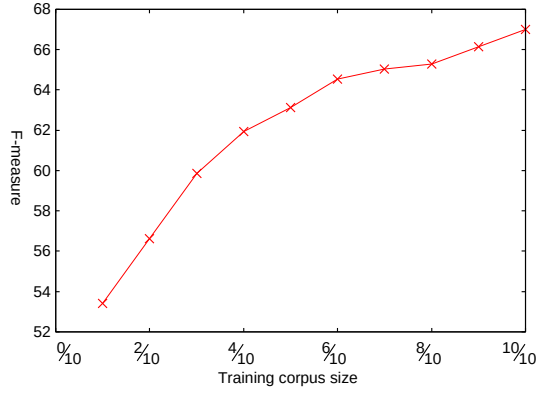


Step 2. 固有表現抽出

初期モデル: 53.42%

⇓ (5 時間)

資源追加後: 67.02%

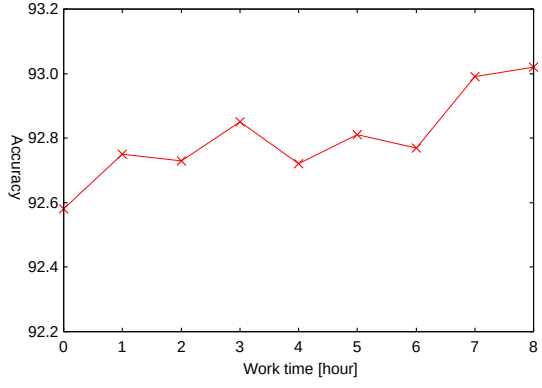


Step 3. 係り受け解析

一般モデル: 92.58%

⇓ (8 時間)

分野適応後: 93.02%



全体の評価

1. 述語項構造 (有向グラフの部分グラフ)
   - 述語と項の組

     例: ⟨ 煮立て, を:水-４００-ｃｃ ⟩, ⟨ 煮立て, で:鍋 ⟩

   - F 値

     初期モデル: 42.01%　多くの研究では辞書追加程度

     ⇓ (8 + 5 + 8 時間)　28.0%のエラーを削減!

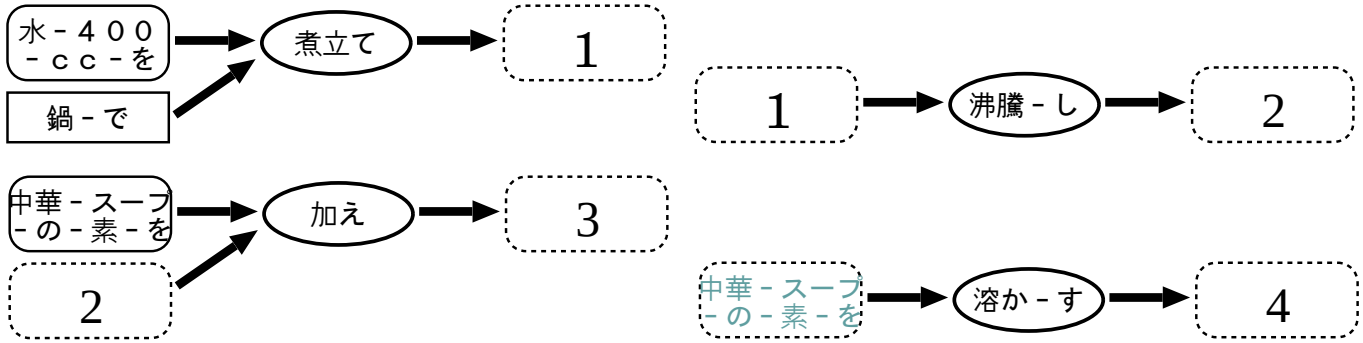
     資源追加後: 58.27%

   - 依然として低い F 値
     - さらなるアノテーション (21 時間 $\ll \infty$)
     - 固有表現認識が問題 (67.02% $\ll$ 90%)
     - それぞれの処理のみを適応した結果を定量的に比較!!

ここから
現在取り組み中

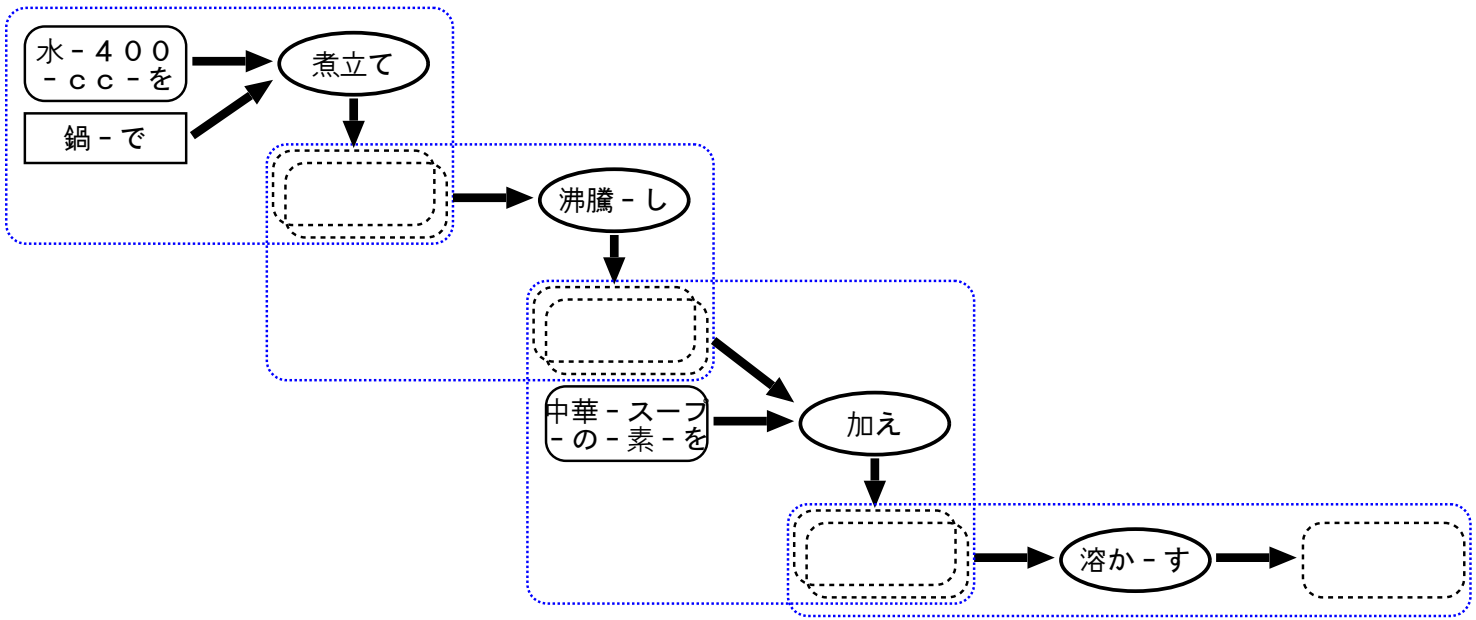
有向グラフへの変換

1. 述語項構造は部分グラフ



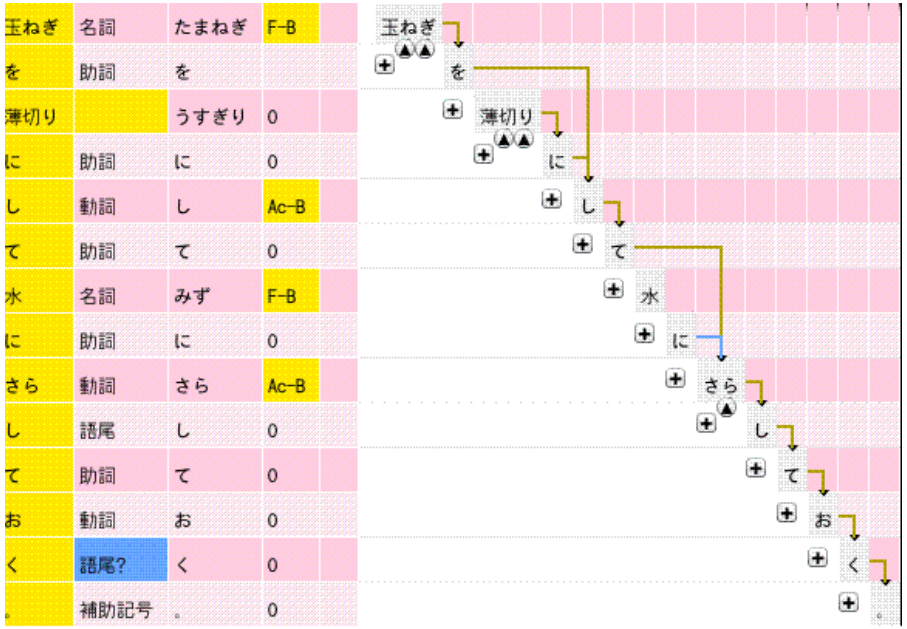
2. 共参照解析による同一固有表現の認識

3. 同一固有表現を同一ノードとして部分グラフを連結



各処理の学習コーパスの充実

1. 単語分割
2. 固有表現認識
3. 係り受け解析
4. 述語項構造解析
5. 共参照解析?



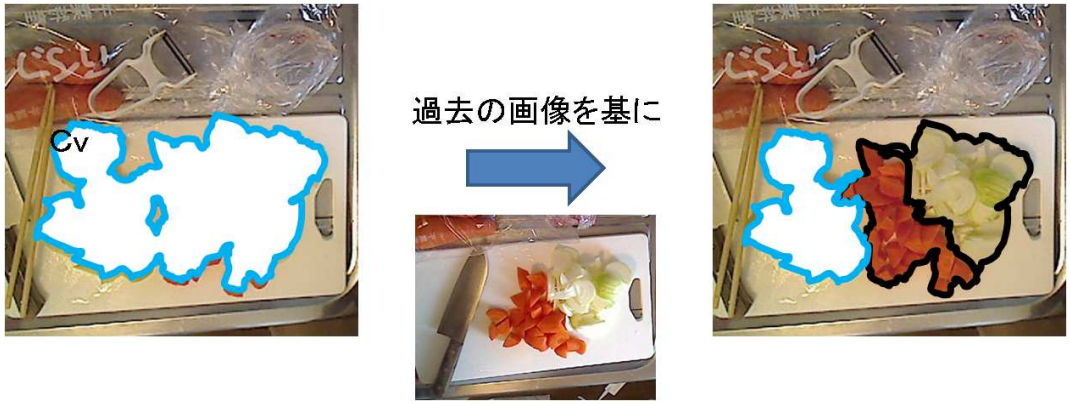
アノテーションツール PNAT (現在 1〜3 のみ対応)

- 各処理の部分的アノテーション大幅増量
  - 部分的アノテーションからの系列予測学習 (≠ 点予測)
- 各処理の改善による全体の精度の定量的評価
  - どの処理のアノテーションに注力?
  - アノテーション or 手法の改善?

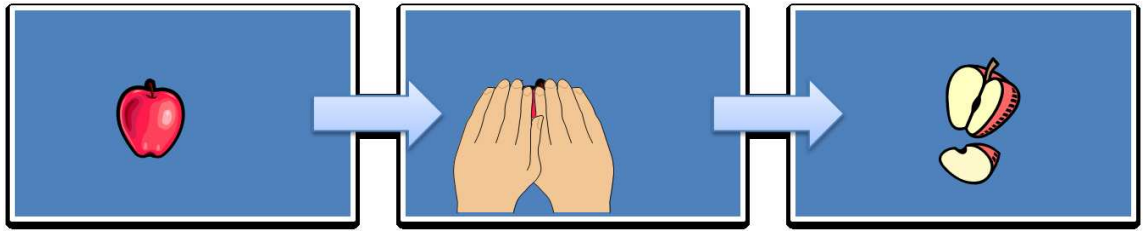
映像処理側の状況

調理映像からの物体ノードの抽出

- １つの連結領域 ≠ １つの物体ノード
  - × 近接して置かれた異種の食材が１つに
  - ⇒ 時間的継続性を考慮し、別々の領域として検出



  - × １つの食材が分割され複数の領域に
  - ⇒ 調理者の手によって加工・移動された領域を追跡
    - 把持の前後で食材領域間を対応付け

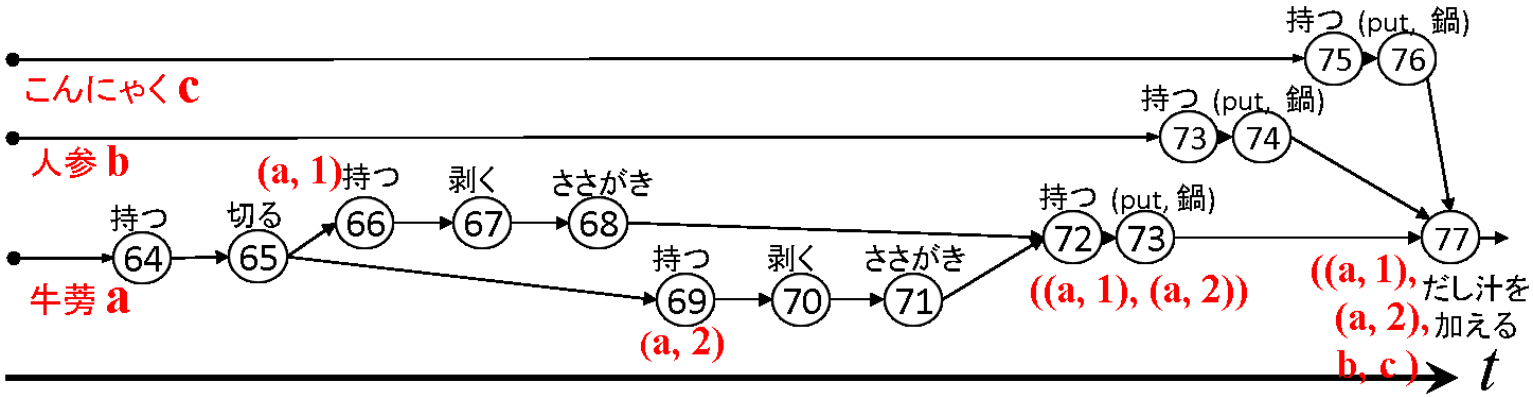


調理映像からの動作ノードの抽出

- 全ての物体は調理者によって加工・移動
  - 物体領域の追跡から動作を推定
    - 物体 (食材・道具) の移動
    - 食材の切削加工
  - 調理者の動きのパターンを検出
    - かき混ぜる: 周期的運動

有向グラフの出力

- 物体と動作の認識
- 時系列情報の利用




```
...
64 (a, 持つ)
65 (a, 切る)
66 ((a,1), 持つ)
67 ((a,1), 皮を剥く)
68 ((a,1), ささがき)
```

```
69 ((a, 2), 持つ)
70 ((a, 2), 皮を剥く)
71 ((a, 2), ささがき)
72 (((a, 1), (a, 2)), 持つ)
73 (((a, 1), (a, 2)), (put, 鍋))
74 (b, 持つ)
```

```
75 (b, (put, 鍋))
76 (c, 持つ)
77 (c, (put, 鍋))
78 (((a, 1),(a, 2), b, c),
    だし汁を加える)
...
```

言語処理と映像処理の統合

- レシピグラフと調理グラフは大域的構造において類似
  - 言語・映像の最適解の有向グラフをマッチ

  or 確率付き出力の確率を考慮したマッチ

- 言語処理 $\rightleftarrows$ 映像処理
  - 物体・動作のラベル
  - 同一物体の表記揺れ
  - part-of, is-a 関係の推定

レシピテキストと調理映像からの実世界理解

- レシピテキストの言語処理

処理	設計	論文	十分な精度
単語分割	☑	☑	☐
固有表現認識	☑	☐	☐
係り受け解析	☑	☑	☐
述語項構造解析	☑	☐	☐

- 調理シーンの映像処理との統合
- 応用
  - 対話システムによる調理の教示
  - 見本となる調理映像の生成

References

- Flannery, D., Miyao, Y., Neubig, G., and Mori, S.: Training Dependency Parsers from Partially Annotated Corpora, in *Proceedings of the Fifth International Joint Conference on Natural Language Processing* (2011)
- Hamada, R., Ide, I., Sakai, S., and Tanaka, H.: Structural Analysis of Cooking Preparation Steps in Japanese, in *Proceedings of the fifth international workshop on Information retrieval with Asian languages*, No. 8 in IRAL '00, pp. 157–164 (2000)
- Momouchi, Y.: Control Structures for Actions in Procedural Texts and PT-Chart, in *Proceedings of the Eighth International Conference on Computational Linguistics*, pp. 108–114 (1980)

- Mori, S. and Nagao, M.: Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis, in *Proceedings of the 16th International Conference on Computational Linguistics* (1996)
- Neubig, G., Nakata, Y., and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (2011)
- 橋本 敦史, 大岩 美野, 舩冨 卓哉, 上田 真由美, 角所 考, 美濃 導彦■調理行動モデル化のための調理観測映像へのアノテーション, 第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009) (2009)

- 山肩 洋子, 角所 考, 美濃 導彦■調理コンテンツの自動作成のためのレシピテキストと調理観測映像の対応付け, *Transactions of ???*, Vol. J90-DII, No. 10, pp. 2817–2829 (2007)
- 森 信介■自然言語処理における分野適応, 人工知能学会誌, Vol. 27, No. 4 (2012)
- 前川 喜久雄■代表性を有する大規模日本語書き言葉コーパスの構築, 人工知能学会誌, Vol. 24, No. 5, pp. 616–622 (2009)