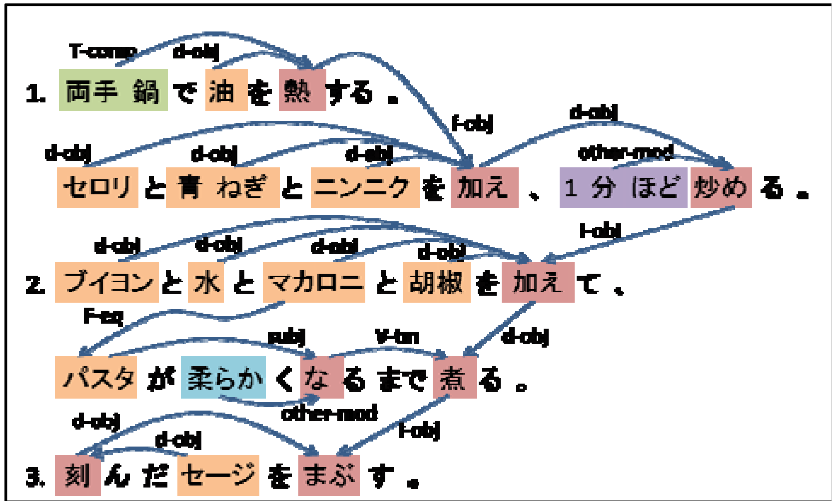
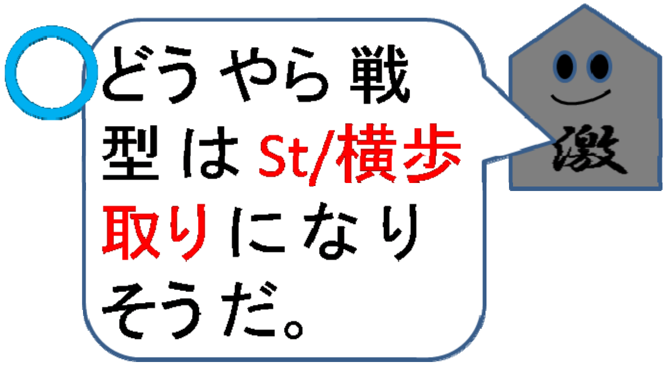
自然言語処理による
テキストの理解に向けて

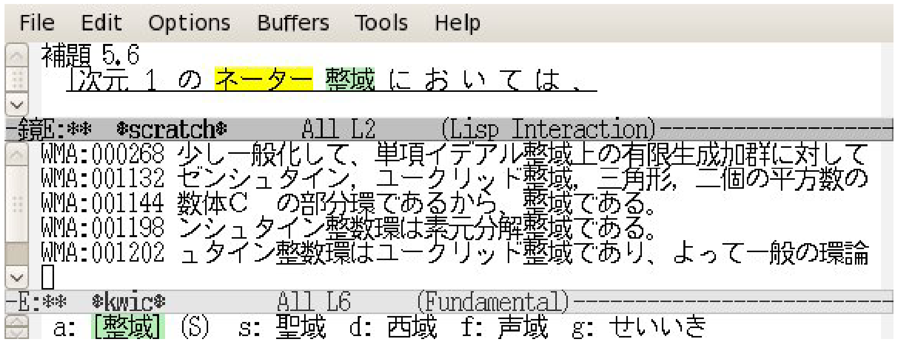
森 信介
京都大学

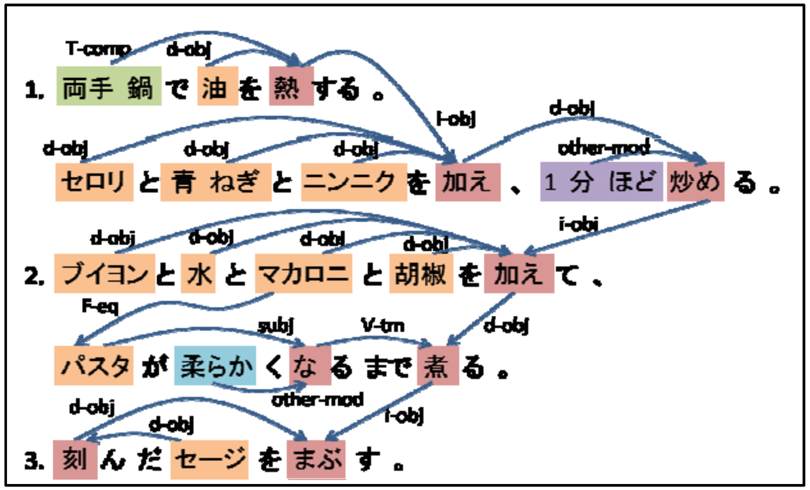


自己紹介

日本語, Englis., França.., Itali..., Esp...., Po.......





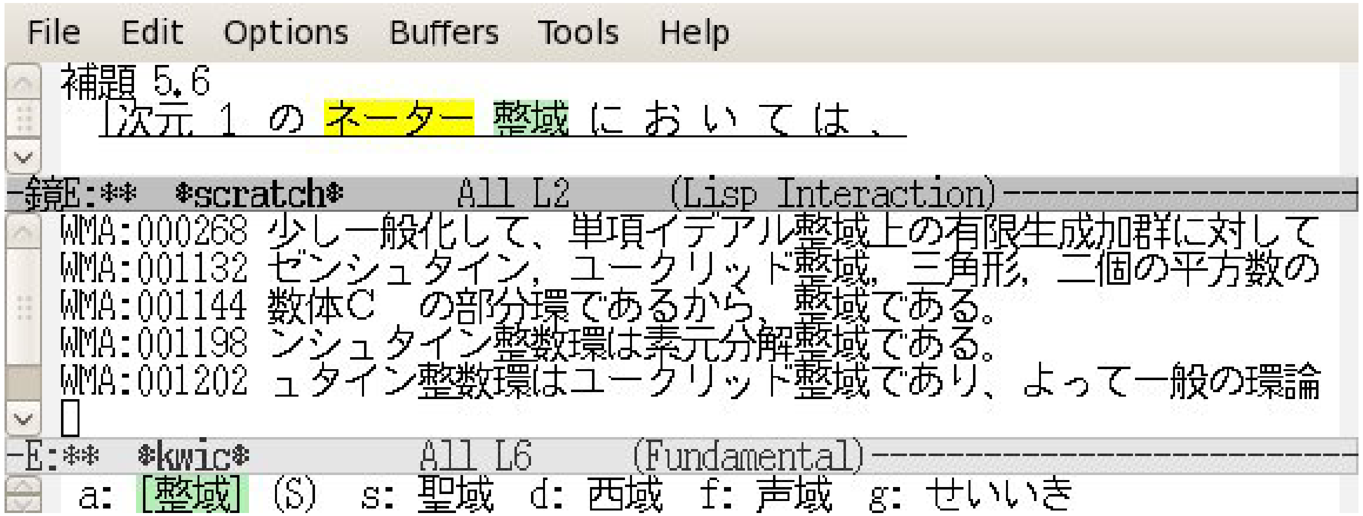


研究経歴

- 京都大学 学士 → 修士 → 博士
  – 自然言語処理
- 日本アイ・ビー・エム
  – 音声認識・合成
- 京都大学学術情報メディアセンター 准教授
  – メディア融合!?

統計的仮名漢字変換

- **確率的モデルによる仮名漢字変換**
  - 森 信介, 土屋 雅稔, 山地 治, 長尾 真
  - 情報処理学会論文誌, Vol.40, No.7, 1999.



- **自作仮名漢字変換を使っている!!**

現在の自然言語処理

形態素解析 →[1] 用語認識 →[2] 構文解析 →[3] 共参照解析 →[4]

1. 彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞
2. 彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞
3. 彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞
4. 彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞

= 小浜バラク　**= 音羽山清水寺**

形態素解析とは

- 形態素解析
  1. 単語分割
  2. 品詞推定
  3. その他タグ (読み, 原型, 正規表記, ...)

INPUT: 吾輩は猫である。
OUTPUT:

我輩	ワガハイ	ワガハイ	我が輩	代名詞		
は	ワ	ハ	は	助詞-系助詞		
猫	ネコ	ネコ	猫	名詞-普通名詞-一般		
で	デ	ダ	助動詞	助動詞-ダ	連用形-一般	
ある	アル	アル	有る	動詞-非自立可能	五段- ラ行	終止形-一般
。	。			補助記号-句点		

EOS

隠れセミマルコフモデルに基づく教師なし完全形態素解析
内海 慶, 塚原 裕史, 持橋 大地, NLP2015

固有表現認識とは

- 固有表現: 単語列とその種類

  種類: 人名, 地名, 組織名, 金額, 日付, 時間, 割合, 人工物
  - 実世界に対応するのは単語列 (単語の定義の問題)
  - 拡張固有表現: 約200種

彼/名詞	Other
と/助詞	Other
清水/名詞	Loc-Begin
寺/名詞	Loc-Intermediate
に/助詞	Other
行く/動詞	Other

  - 種類の定義は課題依存

自前のツール

形態素解析

KyTea (きゅーてぃー): 単語分割・品詞推定・読み推定
短単位 (国立国語研究所)

用語認識

PWNER: 学習コーパスが必要 (一般的固有表現以外の場合)

構文解析

EDA Parser: 単語単位の係り受け (≠ 文節)

共参照解析

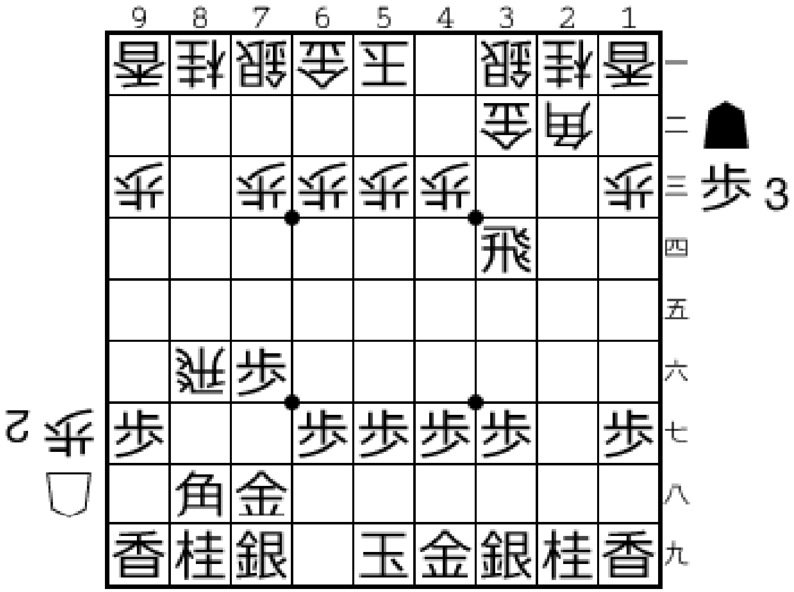
Wikification (後述) を開発中

研究紹介

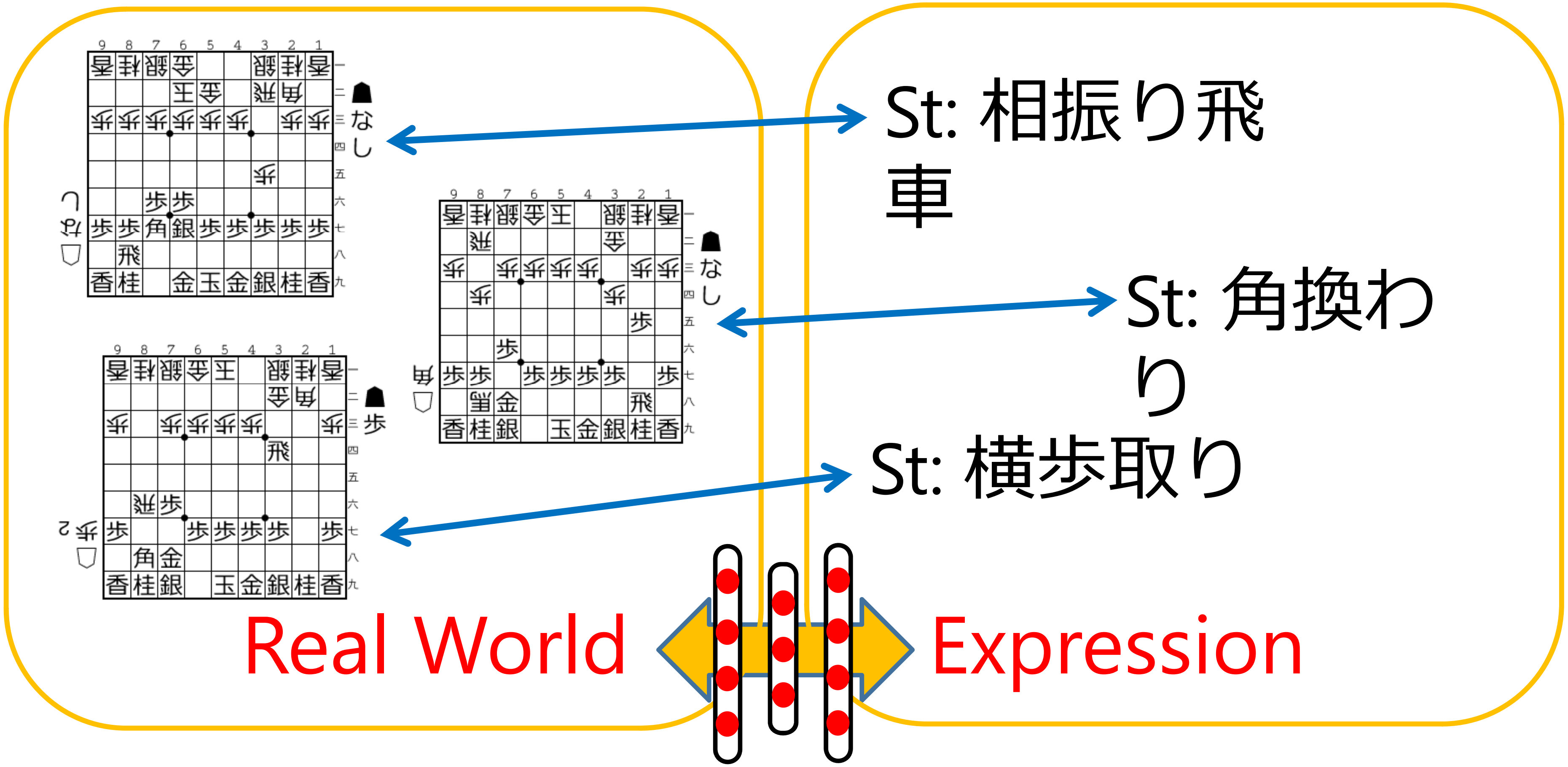
将棋の自動解説

入力: 将棋の局面

出力: 解説



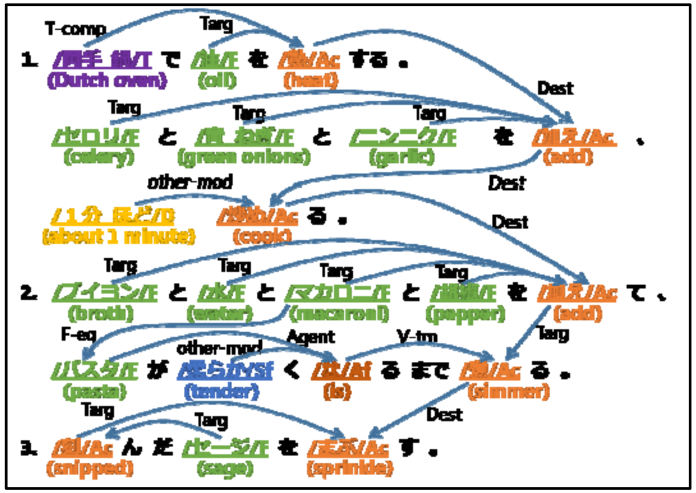
局面と用語の自動対応



- Three layer perceptron
- About 300,000 positions with commentaries
  (with lots of noisy commentaries ...)

研究紹介 1

A Framework for Procedural Text Understanding



Hirokuni Maeta (Cybozu, Inc.)
Tetsuro Sasada (Kyoto Univ.)
<u>Shinsuke Mori </u>(Kyoto Univ.)

Special thanks to
Prof. Yoko Yamakata
and *Kakenhi* Grant

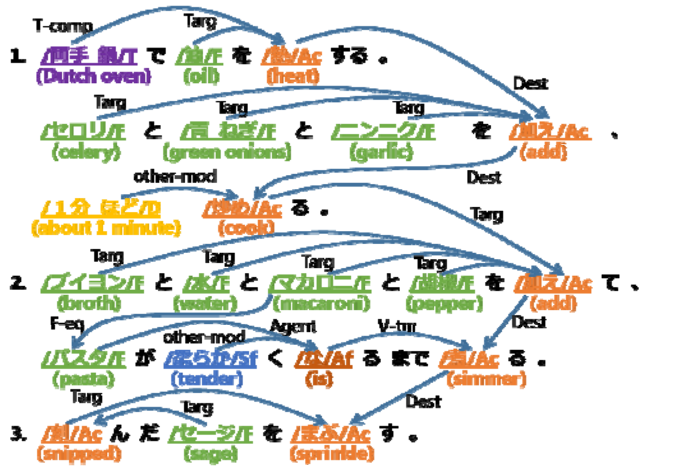
Overview

Procedural text

1. 両手鍋で油を熱する。
   セロリと青ねぎとニンニクを加え、
   １分ほど炒める。
2. ブイヨンと水とマカロニと胡椒を加えて、
   パスタが柔らかくなるまで煮る。
3. 刻んだセージをまぶす。

Today's Talk

Flow graph [Mori+, LREC14]



Cooking robot

[Bollini+, ISER13]

www.denso.co.jp

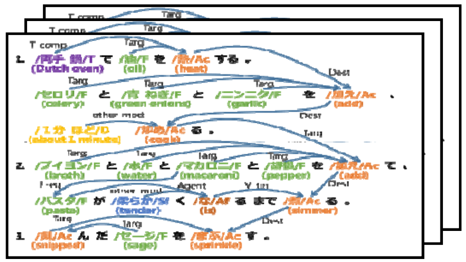
Smart kitchen

[Hashimoto+, IPMU08]

Cut garlic
Then add it

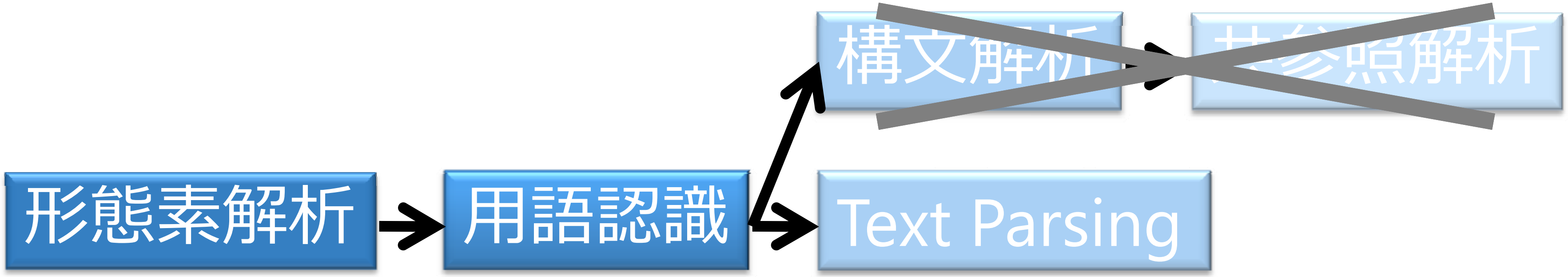
Intelligent search

[Yamakata+, SocInfo13]



設計思想

1. 現在の自然言語処理で自動推定可能



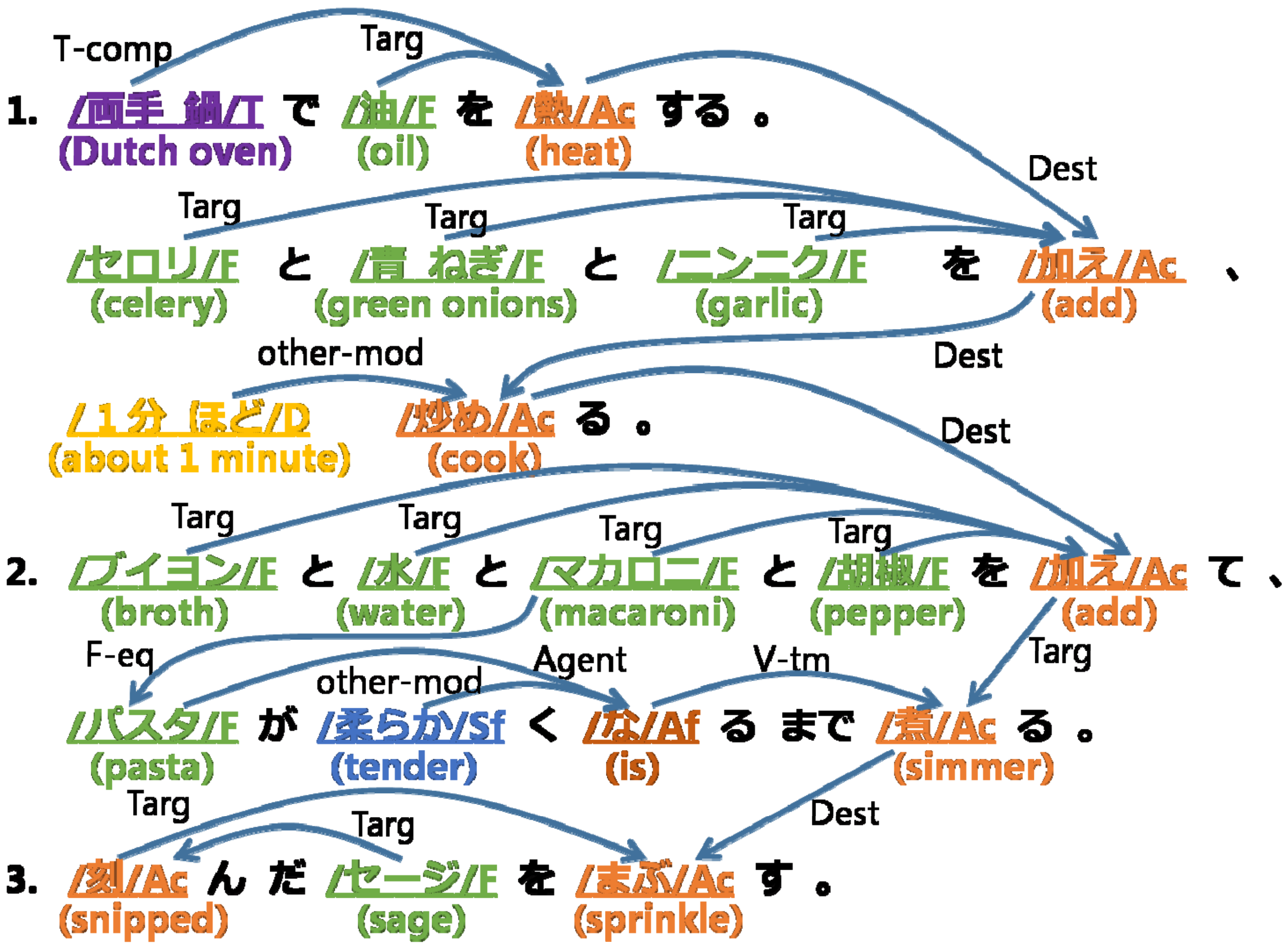
2. アノテーターに明快な基準

3. 世界史についても取り組み開始

FLOW GRAPHS
FOR PROCEDURAL TEXT

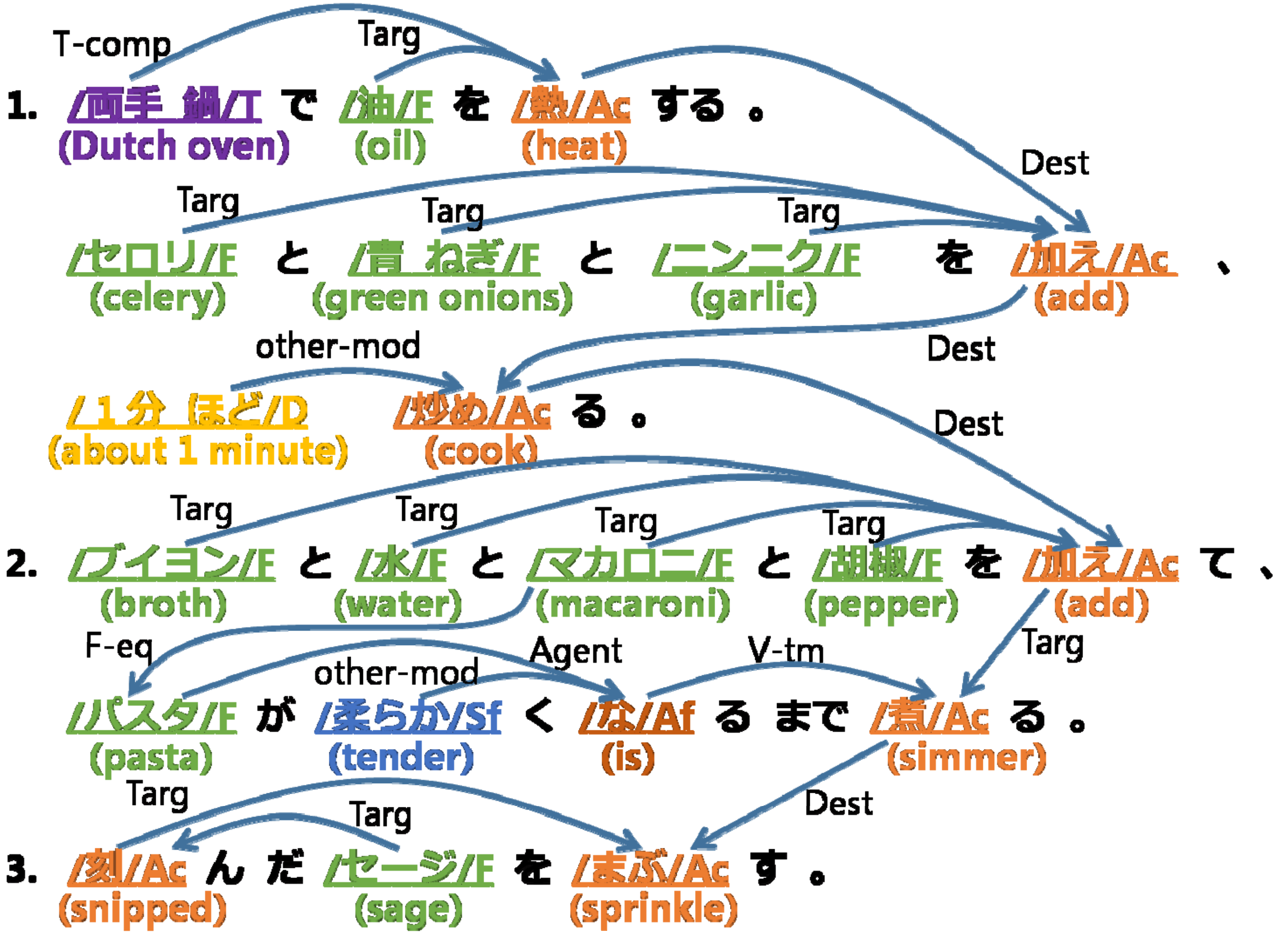
Flow Graph

- **Direct acyclic graph (DAG) representing work flow**



Vertices

- **Concepts = Important terms (word sequence)**

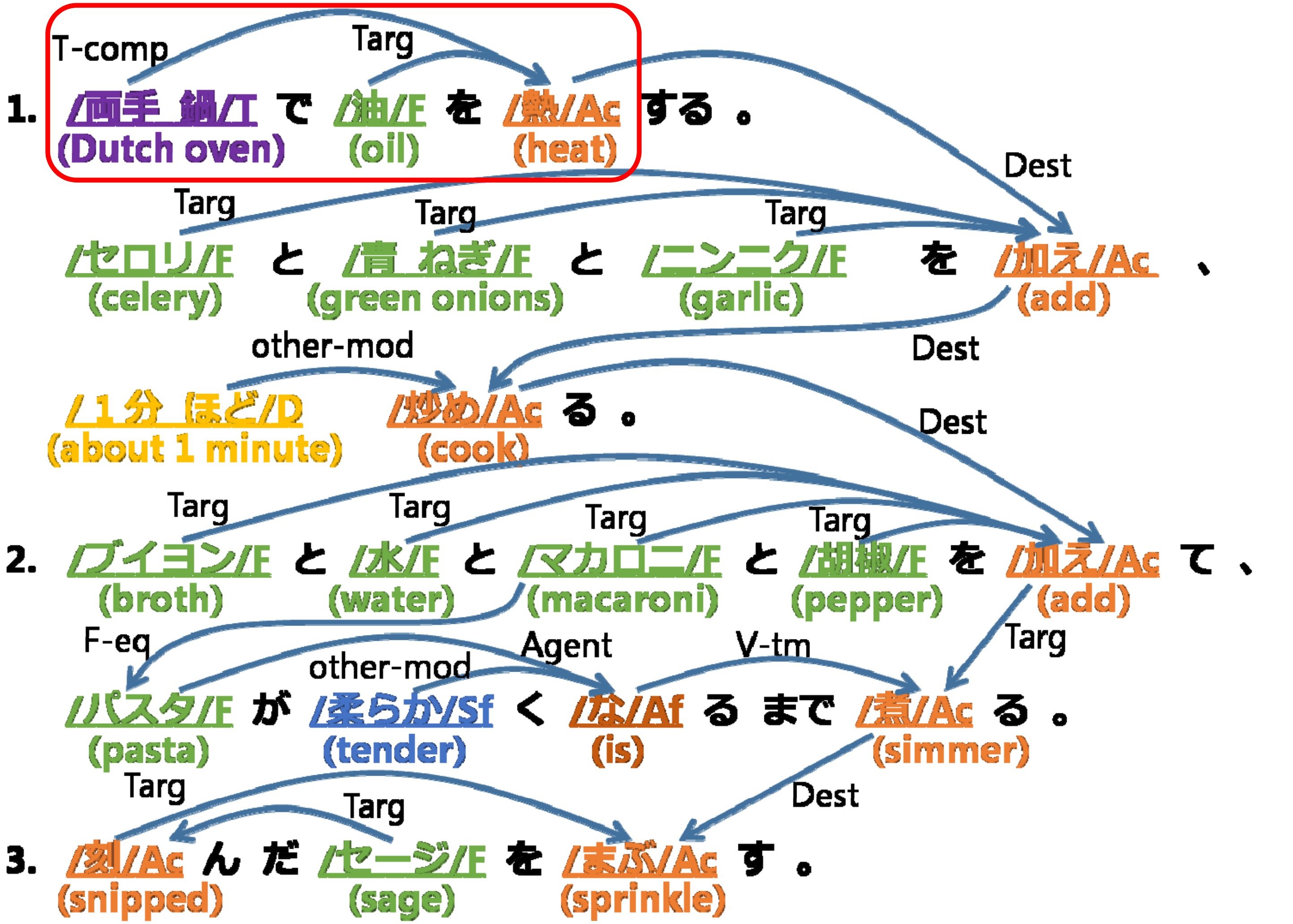


Vertex Labels

Tag	Meaning	Remarks
F	Food	Eatables including intermediate products,
T	Tool	Knife, container, etc.
D	Duration	Duration for cooking
Q	Quantity	Quantity of food
Ac	Action by chef	Verbs representing chef's actions (stem only)
Af	Action by food	Verbs representing food's actions (stem only)
Sf	Food state	Food's initial or intermediate states
St	Tool state	Tool's initial or intermediate states

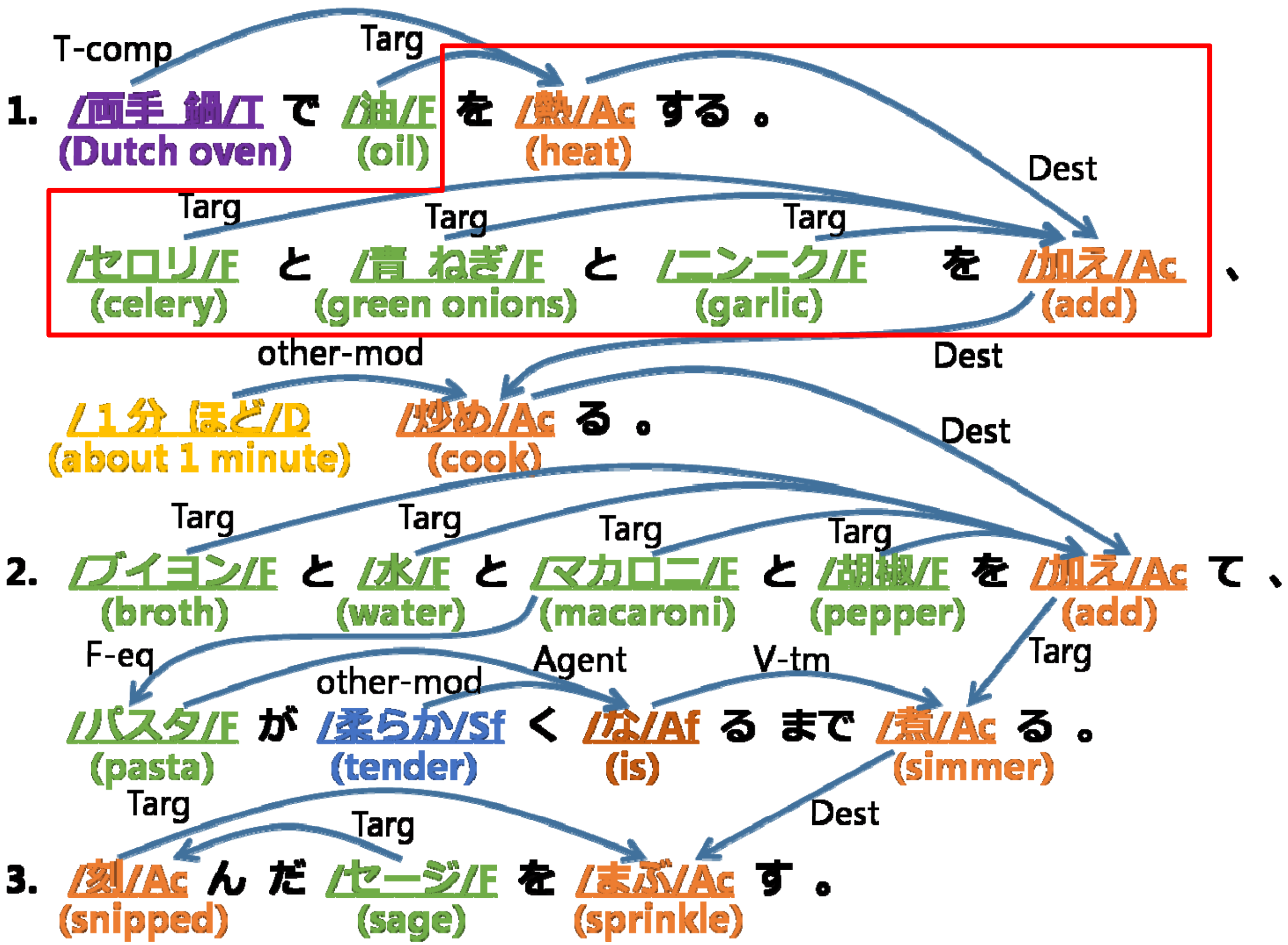
Arcs

- **Relationships between two important terms**



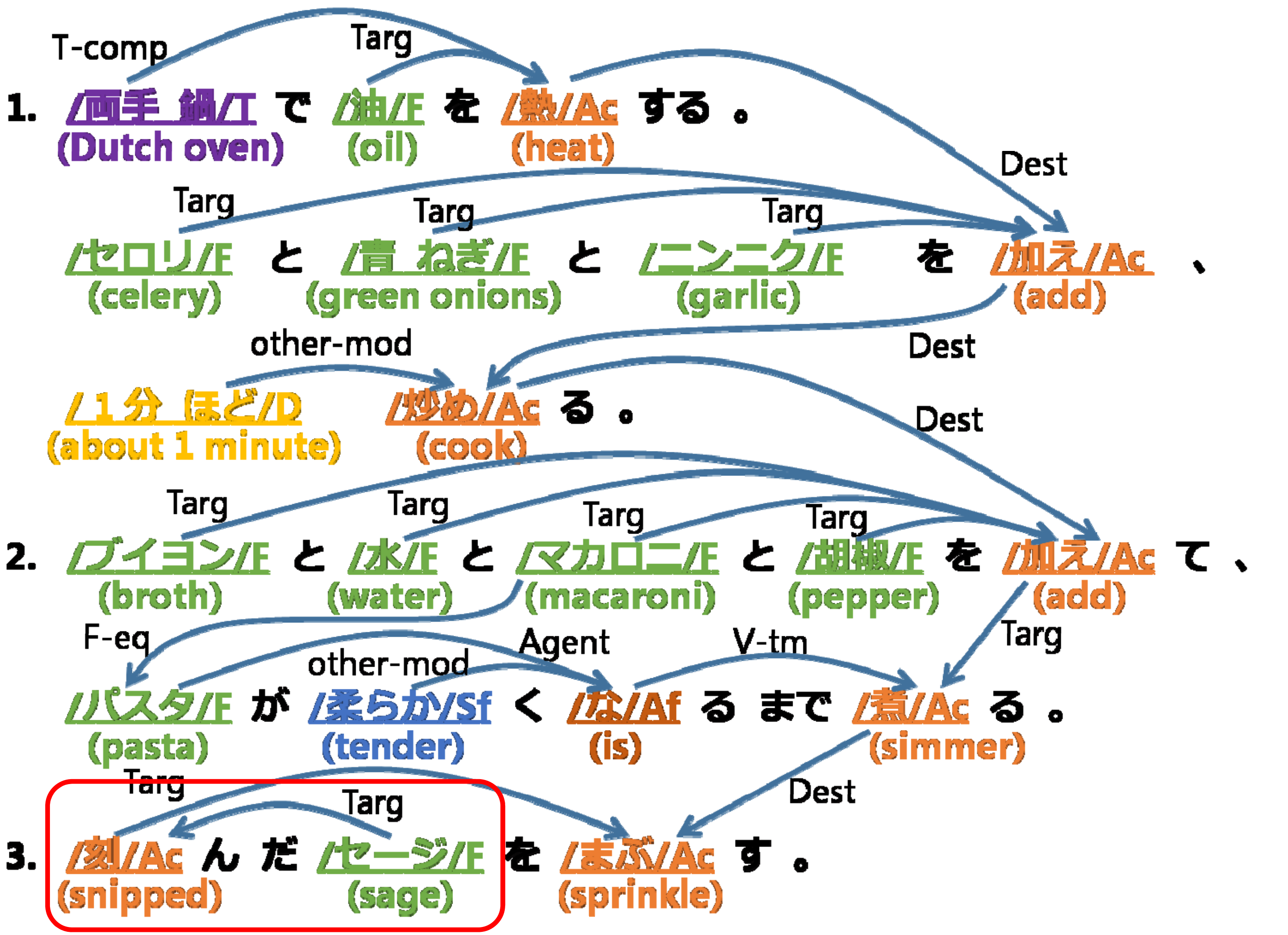
Arcs

- **Relationships between two important terms**



Arcs

- **Relationships between two important terms**



Arc Labels

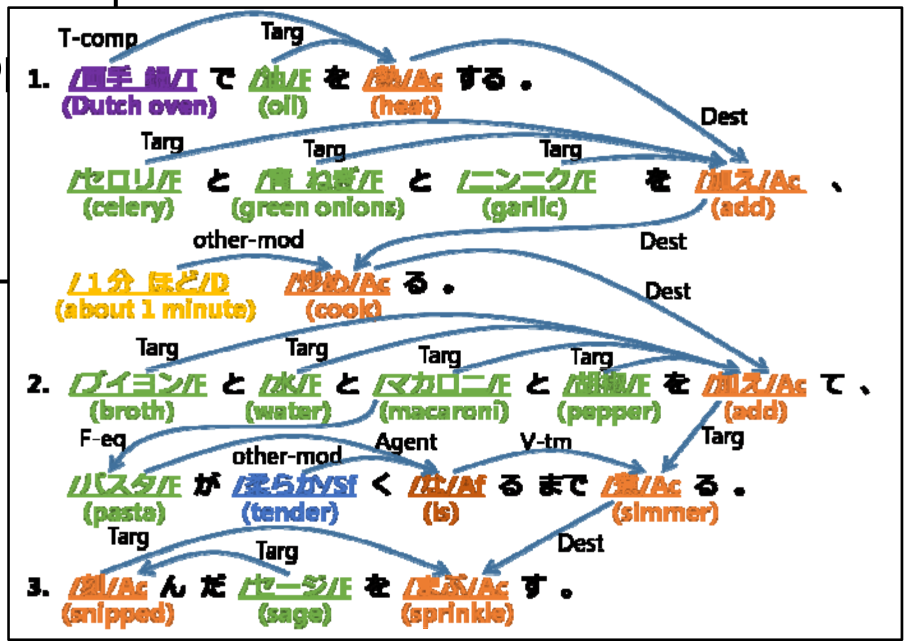
Label	Meaning	Remark
Agent	Action agent	Relationship with actions (Ac or Af)
Targ	Action target	
Dest	Action destination	
T-comp	Tool complement	Tool used in an action
F-comp	Food complement	Food used as a tool
F-eq	Food equality	Identical food
F-part-of	Food part-of	Refer to a part of a food
F-set	Food set	Refer to a set of foods
T-eq	Tool equality	Identical tool
T-part-of	Tool parf-of	Refer to a part of a tool
A-eq	Action equality	Identical action (Ac, Af)
V-tm	Head verb for timing, etc.	
other-mod	Other relationships	

PROCEDURAL TEXT UNDERSTANDING

Procedural Text Understanding

1. 両手鍋で油を熱する。
   (In a Dutch oven, heat oil.)
   セロリと青ねぎとニンニクを加え、
   (Add celery, green onions, and garlic.)
   1分ほど炒める。
   (Cook for about 1 minute.)
2. ブイヨンと水とマカロニと胡椒を加えて、
   パスタが柔らかくなるまで煮る。
   (Add broth, water, macaroni, and pep
   and simmer until the pasta is tender.)
3. 刻んだセージをまぶす。
   (Sprinkle the snipped sage.)

1. Word segmentation
2. Concept identification
3. Flow graph estimation



Procedural Text

1. **両手鍋で油を熱する。**
   (In a Dutch oven, heat oil.)

   セロリと青ねぎとニンニクを加え、1分ほど炒める。
   (Add celery, green onions, and garlic. Cook for about 1 minute.)

2. **ブイヨンと水とマカロニと胡椒を加えて、**
   (Add broth, water, macaroni, and pepper,

   パスタが柔らかくなるまで煮る。
   and simmer until the pasta is tender.)

3. **刻んだセージをまぶす。**
   (Sprinkle the snipped sage.)

Step 1. Word Segmentation

Result:

1. **両手 鍋 で 油 を 熱 する 。**
   (In a Dutch oven, heat oil.)

   セロリ と 青 ねぎ と ニンニク を 加え 、 1 分 ほど 炒め る 。
   (Add celery, green onions, and garlic. Cook for about 1 minute.)

2. **ブイヨン と 水 と マカロニ と 胡椒 を 加え て 、**
   (Add broth, water, macaroni, and pepper,

   パスタ が 柔らか く な る まで 煮 る 。
   and simmer until the pasta is tender.)

3. **刻 ん だ セージ を まぶ す 。**
   (Sprinkle the snipped sage.)

Step 1. Word Segmentation

- Pointwise solution [Neubig+, ACL11]
  - Binary classification problem at each point
  - Trainable from partially segmented sentences
  - KyTea: http://www.phontron.com/kytea/

両手鍋で油を熱する。

? ? ? ? ? ? ? ? ?

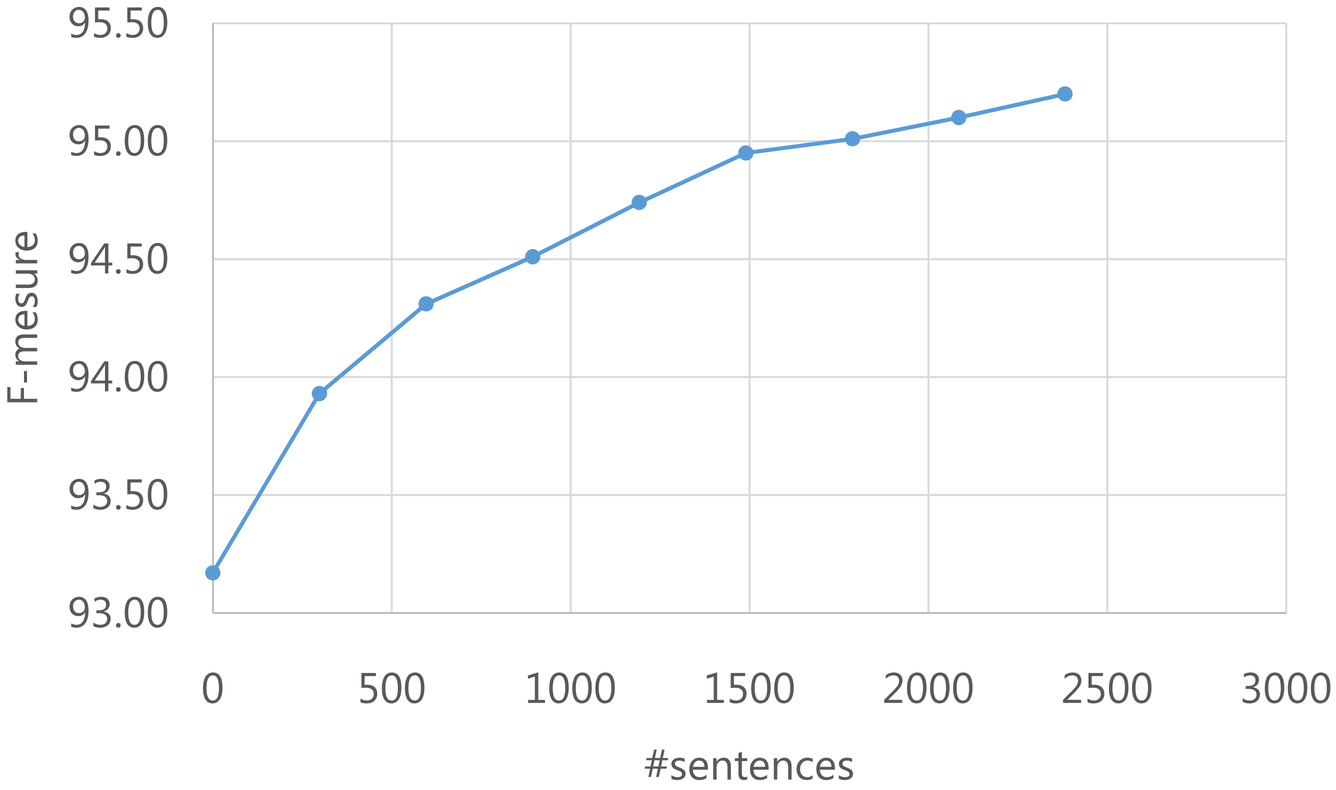
N Y Y Y Y Y Y N Y

両手 鍋 で 油 を 熱 する 。

Step 1. Word Segmentation

- 学習コーパスと異なるテキストに弱い
  - 95%
  - 多くの言語処理応用の研究ではこのまま
- 対象分野の学習コーパスを追加
  - 99%
  - How much?

Twitter の場合



- ベースラインはかなり低い
  - やっぱり Twitter の形態素解析は難しい
- 学習コーパス追加で普通に上がる
  - エラー分析しなくても言語資源追加で簡単に上がる

Step 2. Concept Identification

Result:

1. /両手　鍋/T で /油/F を /熱/Ac する 。
(Dutch oven) (oil) (heat)

/セロリ/F と /青　ねぎ/F と /ニンニク/F を /加え/Ac 、
(celery) (green onions) (garlic) (add)

/１分　ほど/D /炒め/Ac る 。
(about 1 minute) (cook)

2. /ブイヨン/F と /水/F と /マカロニ/F と /胡椒/F を /加え/Ac て 、
(broth) (water) (macaroni) (pepper) (add)

/パスタ/F が /柔らか/Sf く /な/Af る まで /煮/Ac る 。
(pasta) (tender) (is) (simmer)

3. /刻/Ac ん だ /セージ/F を /まぶ/Ac す 。
(snipped) (sage) (sprinkle)

Step 2. Concept identification

- Same as NER [Mori+, CwC12]
  1. BIO2 tagging for each word with a score
  2. Search for the best sequence
  – PWNER: http://plata.ar.media.kyoto-u.ac.jp/tool/

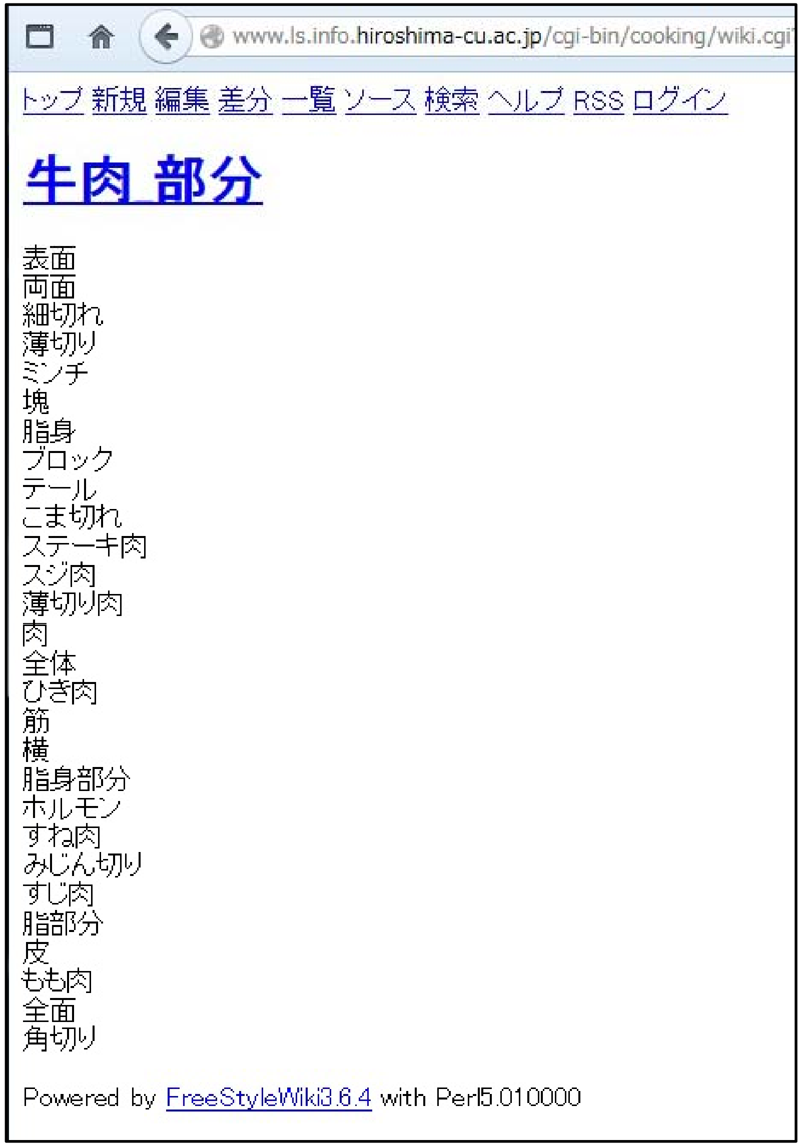
prob.	両手	鍋	で	油	を	熱	する	。
T-B	0.9	0.2	0	0	0	0	0	0
T-I	0.1	0.8	0	0	0	0	0	0
F-B	0	0	0	0.8	0	0	0	0
…								
Other	0	0	1	0	1	0	0	0

Step 2. Concept Identification

- 固有表現認識の場合
  - 人名, 地名, 組織名, 金額, 日付, 時間, 割合, 人工物
  - 10,000文の学習データ
  - 90%
- レシピの場合
  - 分野特有の定義
  - 3,000文の学習データ
  - 90%
  - How much?

Cf. 料理オントロジー構築プロジェクト

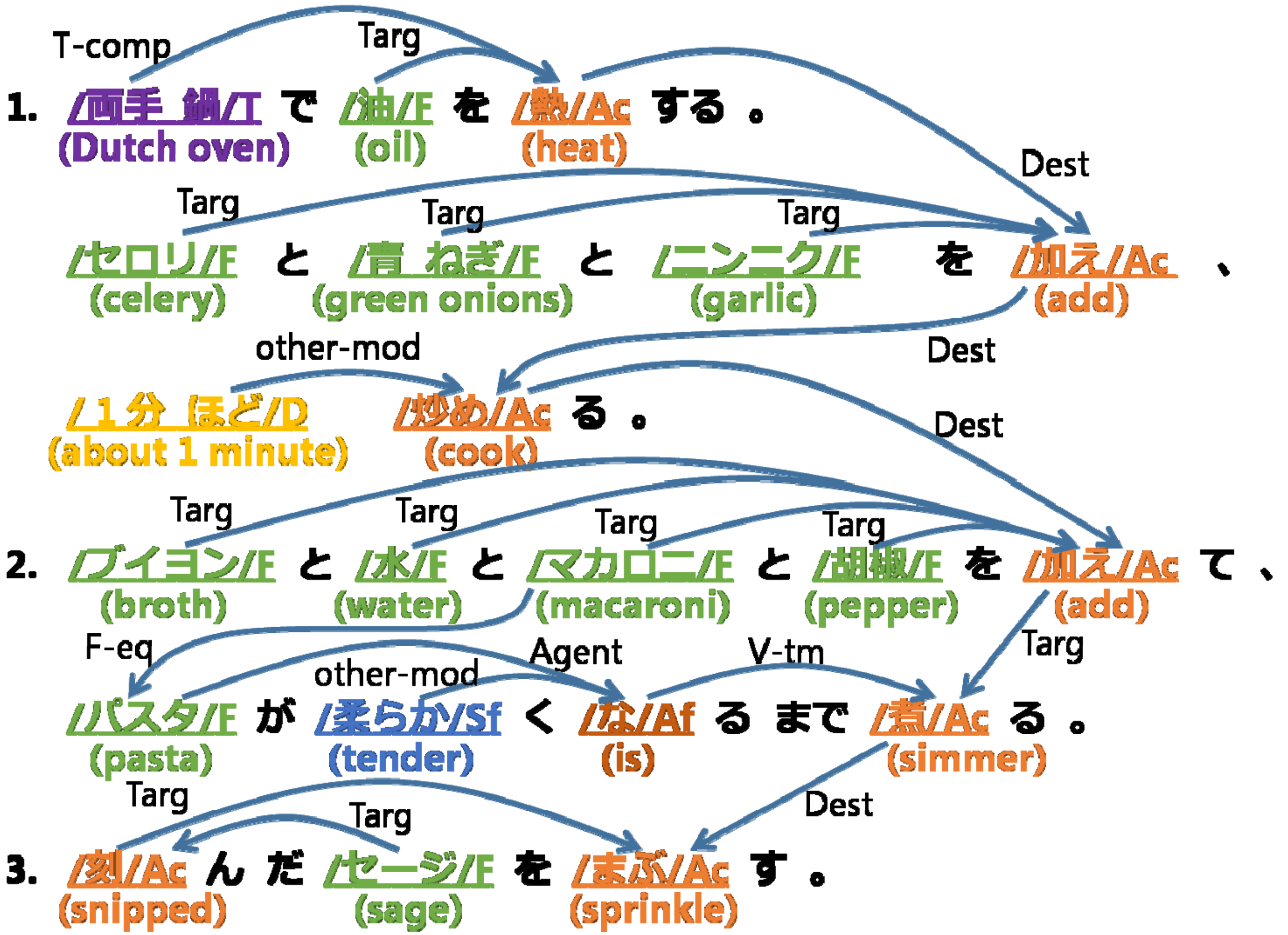
- レシピ用語の階層関係
- 難波先生 (広島市立大)
- 単位が我々と共通



Step 3. Flow Graph Estimation

Parsing of a document, not a sentence

Result:



Step 3. Flow Graph Estimation

1. Maximum Spanning Tree search

- Logistic regression (LR) scores
- 1^{st} order

2. Arc addition

- DAG constrains
- LR score ⋚ penalty(n)

Step 3-1. Tree Estimation

MST using scores by Logistic Regression

2. /ブイヨン/F と /水/F と /マカロニ/F と /胡椒/F を /加え/Ac て 、
/パスタ/F が /柔らか/Sf く /な/Af る まで /煮/Ac る 。

LR score + MST

Targ Targ

2. /ブイヨン/F と /水/F と /マカロニ/F と /胡椒/F を /加え/Ac て 、

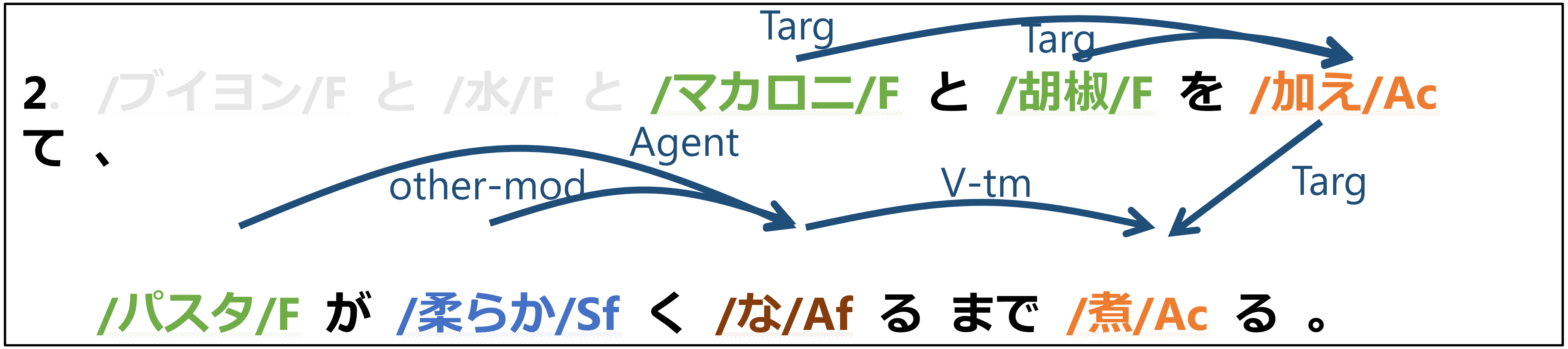
Agent

other-mod V-tm Targ

/パスタ/F が /柔らか/Sf く /な/Af る まで /煮/Ac る 。

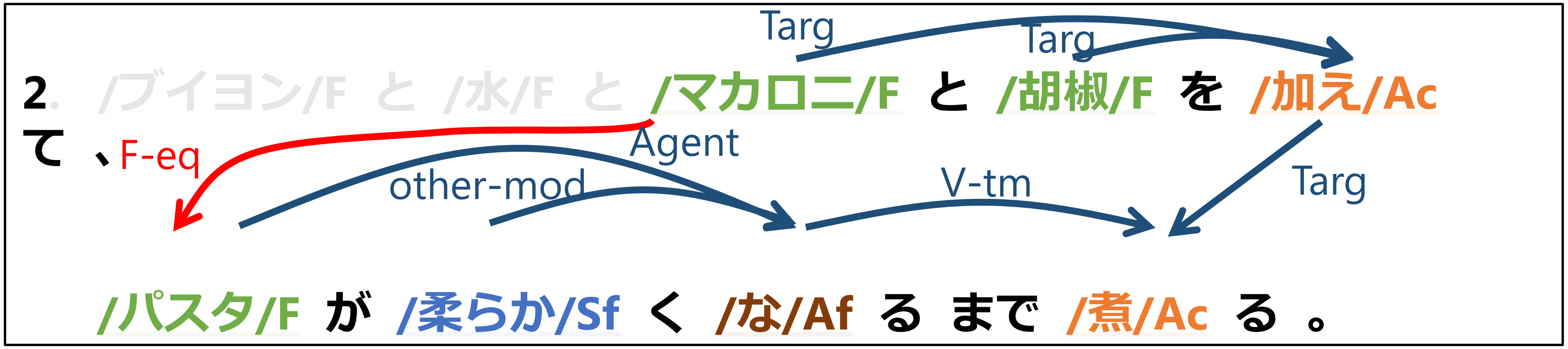
Step 3-2. Arc Addition

Arc addition under some constraints



LR score > $\xi / \lambda e^{-\lambda n}$

Penalty for *n*-th arc addition



Features of a Labeled Arc

u
... で /油/F を /熱/Ac する 。
(... heat oil.)

Dest

v
... と /ニンニク/F を /加え/Ac 、 /１分 ほど/D /炒め/Ac る 。
(Add ... and garlic. Cook for about 1 minute.)

Combination of the followings:
- 3 surrounding words of ***u*** or ***v***
- function words between ***u*** and ***v***
- arc label
- word sequence, pronunciation, and concept tag of ***u*** or ***v***
- whether ***u*** and ***v*** in the same sentence or not
- whether ***u*** and ***v*** in the same step or not
- whether an Ac exists between ***u*** and ***v*** or not

Naïve MST Parser Application

- Only concepts are considered.

1. /両手 鍋/T で /油/F を /熱/Ac する 。
   (Dutch oven) (oil) (heat)

   /セロリ/F と /青 ねぎ/F と /ニンニク/F を /加え/Ac 、
   (celery) (green onions) (garlic) (add)

   / 1 分 ほど/D /炒め/Ac る 。
   (about 1 minute) (cook)

2. /ブイヨン/F と /水/F と /マカロニ/F と /胡椒/F を /加え/Ac て 、
   (broth) (water) (macaroni) (pepper) (add)

   /パスタ/F が /柔らか/Sf く /な/Af る まで /煮/Ac る 。
   (pasta) (tender) (is) (simmer)

3. /刻/Ac ん だ /セージ/F を /まぶ/Ac す 。
   (snipped) (sage) (sprinkle)

Procedural Text Understanding

- Words not covered by concepts are clues.

1. /両手 鍋/T で /油/F を /熱/Ac する 。
   (Dutch oven) (oil) (heat)

   /セロリ/F と /青 ねぎ/F と /ニンニク/F を /加え/Ac 、
   (celery) (green onions) (garlic) (add)

   /１分 ほど/D /炒め/Ac る 。
   (about 1 minute) (cook)

2. /ブイヨン/F と /水/F と /マカロニ/F と /胡椒/F を /加え/Ac て 、
   (broth) (water) (macaroni) (pepper) (add)

   /パスタ/F が /柔らか/Sf く /な/Af る まで /煮/Ac る 。
   (pasta) (tender) (is) (simmer)

3. /刻/Ac ん だ /セージ/F を /まぶ/Ac す 。
   (snipped) (sage) (sprinkle)

EVALUATION

Experimental Settings

- Criterion: F-measure
  - Harmonic mean of precision and recall
  - Unit: labeled arcs
- Recipe flow graph corpus (r-FG corpus)
  - Randomly downloaded from CookPad
  - Manually annotated with flow graphs
  - Available at http://plata.ar.media.kyoto-u.ac.jp/data/recipe/

#Recipes	#Sent.	#Vertices	#Words
200	1,303	7,268	25,446

→ 10-fold cross validation

Flow Graph Estimation

- Baseline: naïve application of an MST parser
  – Input: a concept sequence (cf. word sequences)
  – MST + arc addition
  – Ignore words not covered by the concepts

Method	Precision	Recall	F-measure
Baseline	65.1	61.5	63.2
Proposed	73.5	69.1	**71.2**

- Words other than concepts are important for flow graph estimation.

Text Parsing on a Raw Text

Task	Input	F
Word segmentation (WS)	Raw Recipe Texts	98.6
Concept Identification (CI)	Gold WS results	90.7
Flow Graph Estimation (FGE)	Gold WS/CI results	72.1
WS + CI + FGE	Raw Recipe Texts	51.6

- WS
  - Enough high and well solved
- CI
  - Less accurate than WS
  - Needs improvement or more resource
- FGE
  - Least accurate, most difficult
  - Needs more resource or more sophisticated techniques

CONCLUDING REMARKS

Conclusion

Procedural text

1. 両手鍋で油を熱する。
   セロリと青ねぎとニンニクを加え、
   １分ほど炒める。
2. ブイヨンと水とマカロニと胡椒を加えて、
   パスタが柔らかくなるまで煮る。
3. 刻んだセージをまぶす。

Today's Talk

Flow graph **[Mori+, LREC14]**

- Framework for procedural text understanding
  1. Word segmentation (for Japanese or Chinese)
  2. Concept identification
  3. Flow graph construction (most difficult)
     - Referring to words other than concepts as clues

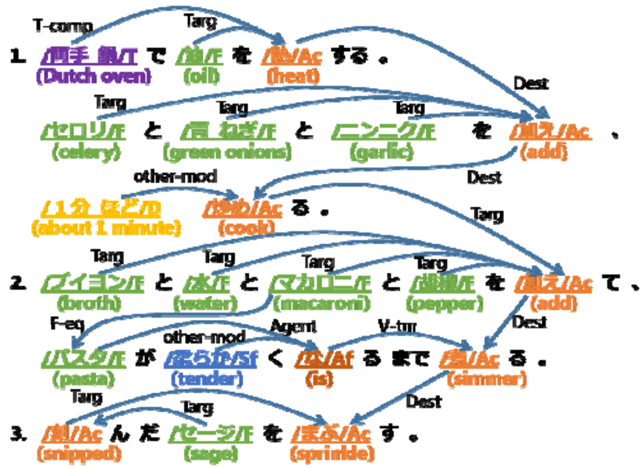
Applications

Procedural text

1. 両手鍋で油を熱する。
   セロリと青ねぎとニンニクを加え、
   １分ほど炒める。
2. ブイヨンと水とマカロニと胡椒を加えて、
   パスタが柔らかくなるまで煮る。
3. 刻んだセージをまぶす。

Today's Talk

Flow graph [Mori+, LREC14]



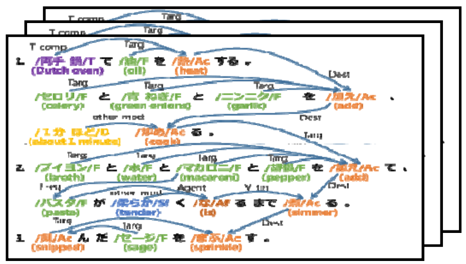
Cooking robot [Bollini+, ISER13]

www.denso.co.jp

Smart kitchen [Hashimoto+, IPMU08]

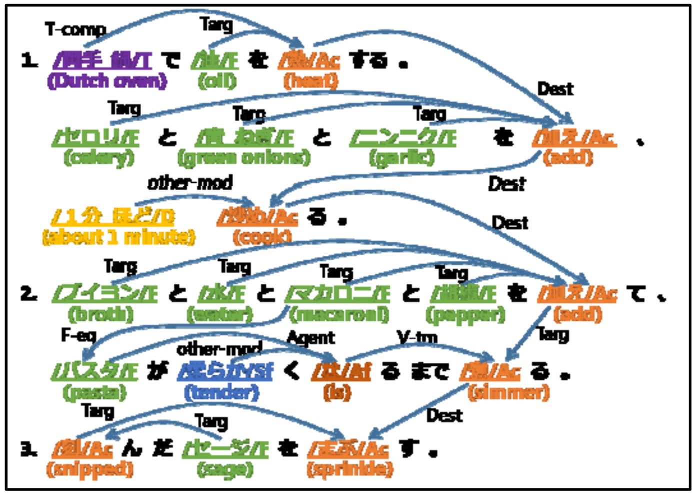
Cut garlic
Then add it

Intelligent search [Yamakata+, SocInfo13]



研究紹介 2

FlowGraph2Text: Automatic Sentence Skeleton Compilation for Procedural Text Generation



Shinsuke Mori (Kyoto Univ.)
et al.

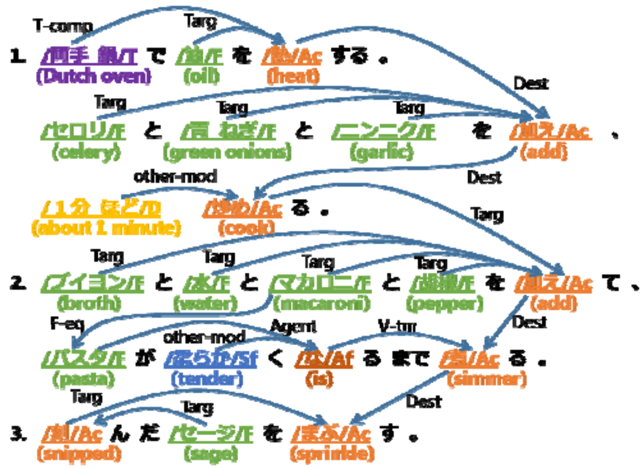
Overview

Procedural text

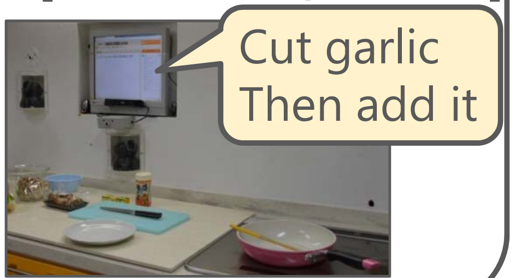
1. 両手鍋で油を熱する。
   セロリと青ねぎとニンニクを加え、
   １分ほど炒める。
2. ブイヨンと水とマカロニと胡椒を加えて、
   パスタが柔らかくなるまで煮る。
3. 刻んだセージをまぶす。

Today's Talk

Flow graph [Mori+, LREC14]

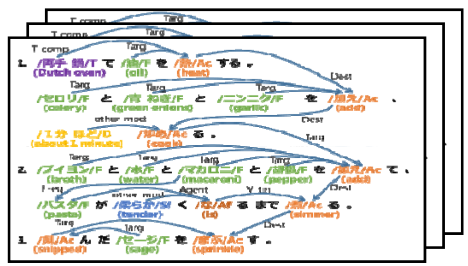


Smart kitchen [Hashimoto+, IPMU08]



Cooking vodeo

Intelligent search [Yamakata+, SocInfo13]



方針

- テンプレートで解決 (文法学習では先が長すぎ)
  - テンプレートを自動生成
- 実際のレシピの重要語を変数化
  - 変数化する箇所を自動認識

概要

1. 部分グラフを抽出

両手 鍋　油　熱　セロリ　青 ねぎ　ニンニク　加え

2. 大量の例文集(Skeleton Pool)の検索

で　を　する　、　、　を　る

3. 内容語を置き換え

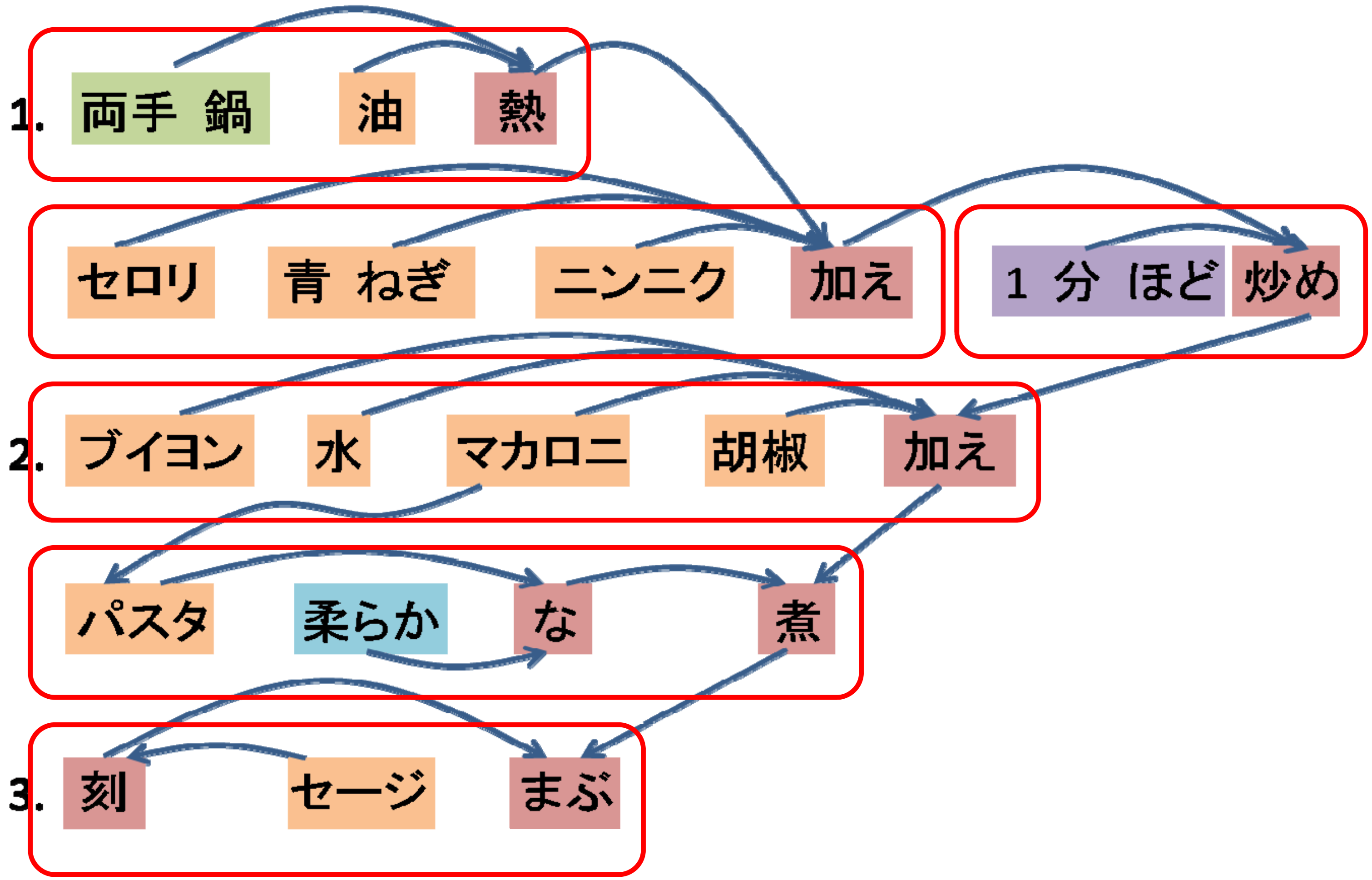
両手 鍋 で 油 を 熱 する	セロリ、青 ねぎ、ニンニク を 加える

研究課題

1. 部分グラフをどう抽出するか？
2. 例文検索の方法は？
3. 置き換えの時の問題点は？
4. どの程度のサイズの例文集が必要か？
5. 生成された文の質は？

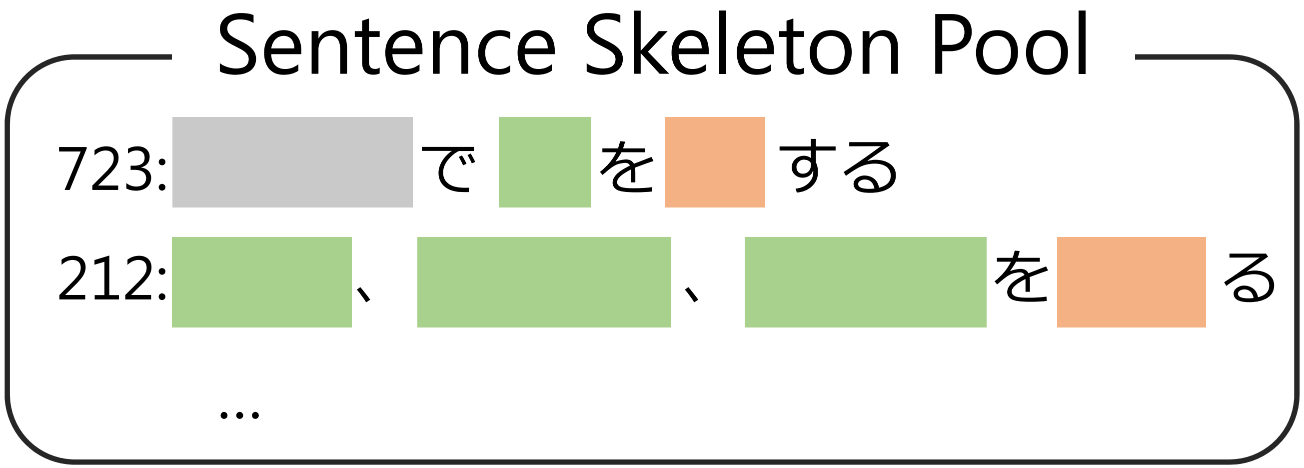
1. 部分グラフの抽出方法

- 調理者動作を含む部分グラフを深さ優先探索



2. 例文検索の方法

1. 大量の例文を収集
2. レシピ用語を自動認識 (精度90%)
3. レシピ用語をクラス名で置き換え



4. 部分グラフのクラス列を検索 (→半順序)
5. 最大頻度の Skeleton を選択 (→言語モデル参照)

研究課題

- ✓ 部分グラフをどう抽出するか？
- ✓ 例文検索の方法は
- ✓ 置き換えの時の問題点は？

4. どの程度のサイズの例文集が必要か？
5. 生成された文の質は？

Skeleton Pool Size を変化させた結果

40レシピ(約300部分木)から文生成

skeleton pool の作成に用いる文の数	2,769 (1/256)	11,077 (1/64)	44,308 (1/16)	177,235 (1/4)	708,940 (1/1)
skeletonなし部分木	52	27	17	9	4
平均 skeleton 数	37.4	124.3	450.2	1598.1	5483.3
BLEU	11.19	11.25	12.86	13.12	13.76

- Skeleton なしの部分木がある
- 平均 skeleton 数は十分 ➡ skeleton 選択を工夫
- BLEU: 元テキストに対する n 単語列のマッチ割合
  - n = 1, 2, 3, 4 の幾何平均

主観評価

Table 5: Result of text quality survey on 10 recipe texts.

	Evaluator 1			Evaluator 2			Evaluator 3			Evaluator 4		
BLEU	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
6.50	13	2	4	11	0	3	12	0	2	7	1	2
7.99	7	2	2	5	2	2	7	1	1	4	2	2
10.09	18	2	4	15	2	1	17	4	1	11	4	2
11.60	24	1	4	13	2	4	18	2	4	13	1	2
13.35	6	1	4	6	0	4	7	1	5	4	1	2
14.70	16	1	4	12	2	4	12	0	3	6	2	2
16.76	9	2	3	6	1	3	7	1	3	5	3	2
19.65	8	2	5	6	1	1	4	1	4	4	2	4
22.85	18	1	4	15	2	5	12	2	2	7	3	2
31.35	5	1	5	5	0	4	5	1	3	5	1	4
Ave.	12.4	1.5	3.9	9.4	1.2	3.1	10.1	1.3	2.8	6.6	2.0	2.4
PCC	−0.30	−0.46	+0.57	−0.24	−0.24	+0.36	−0.46	−0.04	+0.26	−0.29	−0.04	+0.70

PPC stands for Pearson correlation coefficient.

- Q1: 文法誤りの数, Q2: 意味が不明確な表現の数,

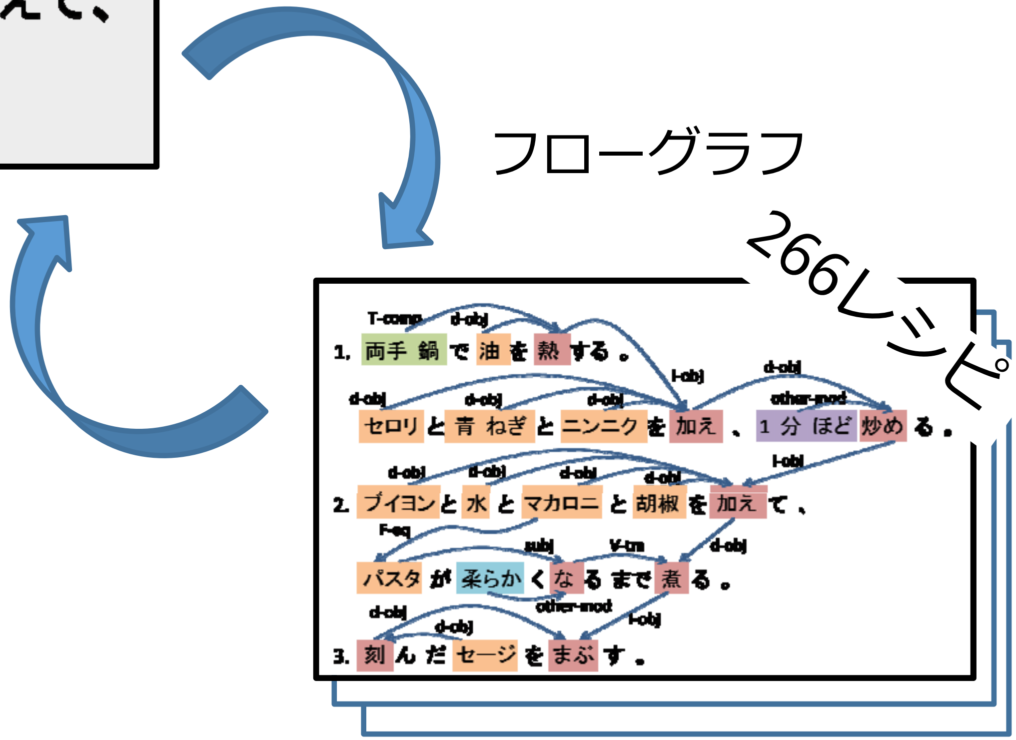
  Q3: 元レシピとの比較 (5段階)

- すべて BLEU と相関がある　機械的評価でOK

テキスト ⇔ グラフ

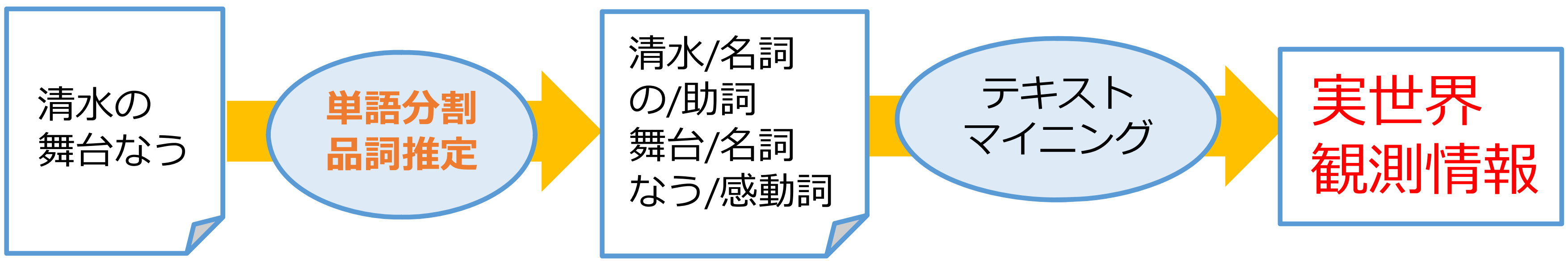
テキスト

1. 両手鍋で油を熱する。
   セロリと青ねぎとニンニクを加え、
   1分ほど炒める。
2. ブイヨンと水とマカロニと胡椒を加えて、
   パスタが柔らかくなるまで煮る。
3. 刻んだセージをまぶす。



インターネット情報の分析

- 市井の声をタイムリーに分析
  - Citizen Sensor [Sheth, IEEE I'net Comp., 2009]
  - マイクロブログの分析 [新田+, DBSJ, 2014]



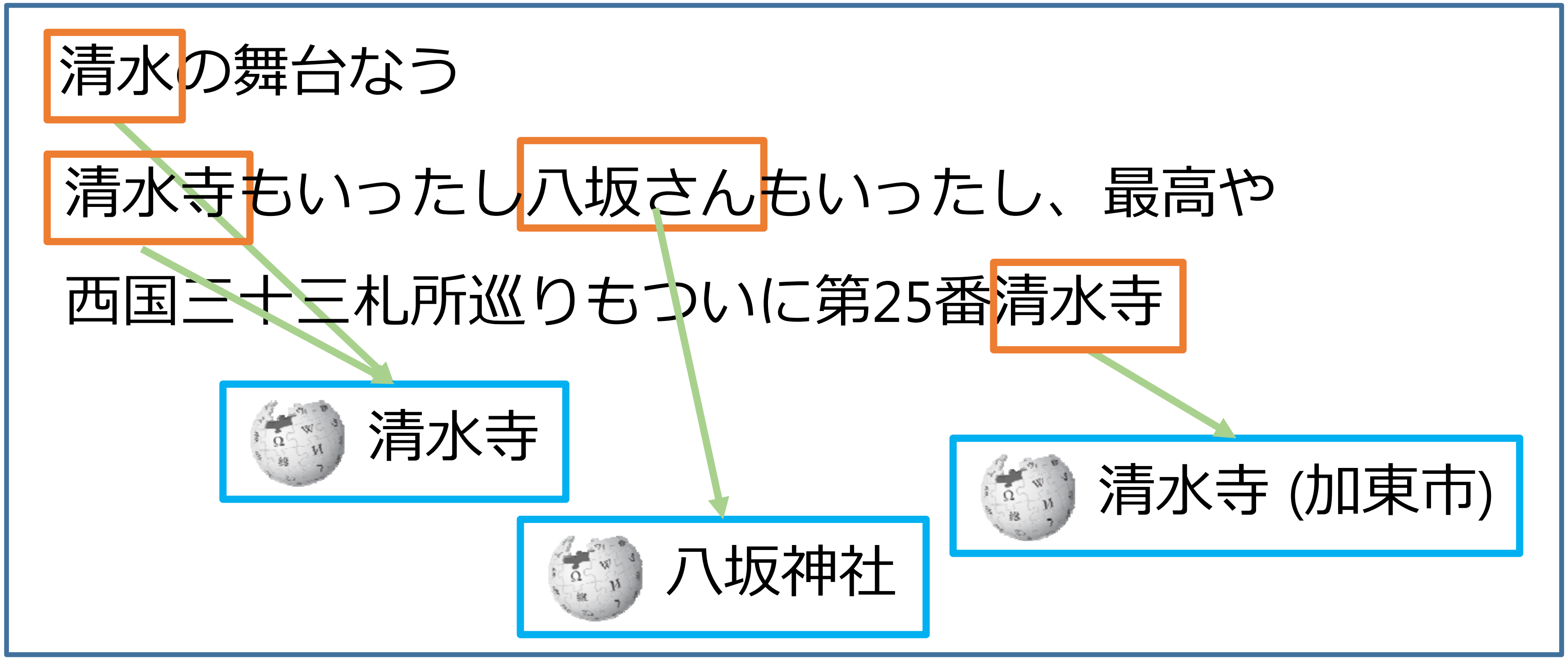
実世界についての自然言語処理

- 言語表現と実世界の対応（固有名詞など）
- 実世界の対象間の関係（寺院／場所など）
  - テキストから Wikipedia のリンク

    Wikification

Wikification

- mention から Wikipedia entry への mapping
  - 表層ゆれ解消+曖昧性解消（多対多）



Wikification の利点

清水寺

座標: 北緯34度59分41.39秒 東経135度47分6.01秒

この記事には**複数の問題があります**。改善やノートページでの議論にご協力ください。

- **出典**がまったく示されていないか不十分です。内容に関する文献や情報源が必要です。（2015年4月）
- 出典は脚注などを用いて記述と関連付けてください。（2015年4月）
- **言葉を濁した**曖昧な記述になっています。（2015年4月）

この項目では、京都市の清水寺について説明しています。
- その他の清水寺については「清水寺 (曖昧さ回避)」をご覧ください。

清水寺（きよみずでら）は、京都府京都市東山区清水にある寺院。山号を音羽山。本尊は千手観音、開基（創立者）は延鎮である。もとは法相宗に属したが、現在は独立して北法相宗大本山を名乗る。西国三十三所観音霊場の第16番札所である。

目次 [非表示]

清水寺

本堂 舞台

所在地	京都府京都市東山区清水一丁目294
位置	北緯34度59分41.39秒 東経135度47分6.01秒

観光ガイド

写真

位置

にない表現の共参照解析

- Wikipediaにない表現
  - 最新の情報
  - エントリーを立てるほどでもない表現

- クラスタリング