文脈情報と
格構造の類似度を用いた
日本語文間述語項構造解析

奈良先端科学技術大学院大学
林部祐太　小町守　松本裕治

第201回自然言語処理研究会
@2011.5.16

本研究の目的

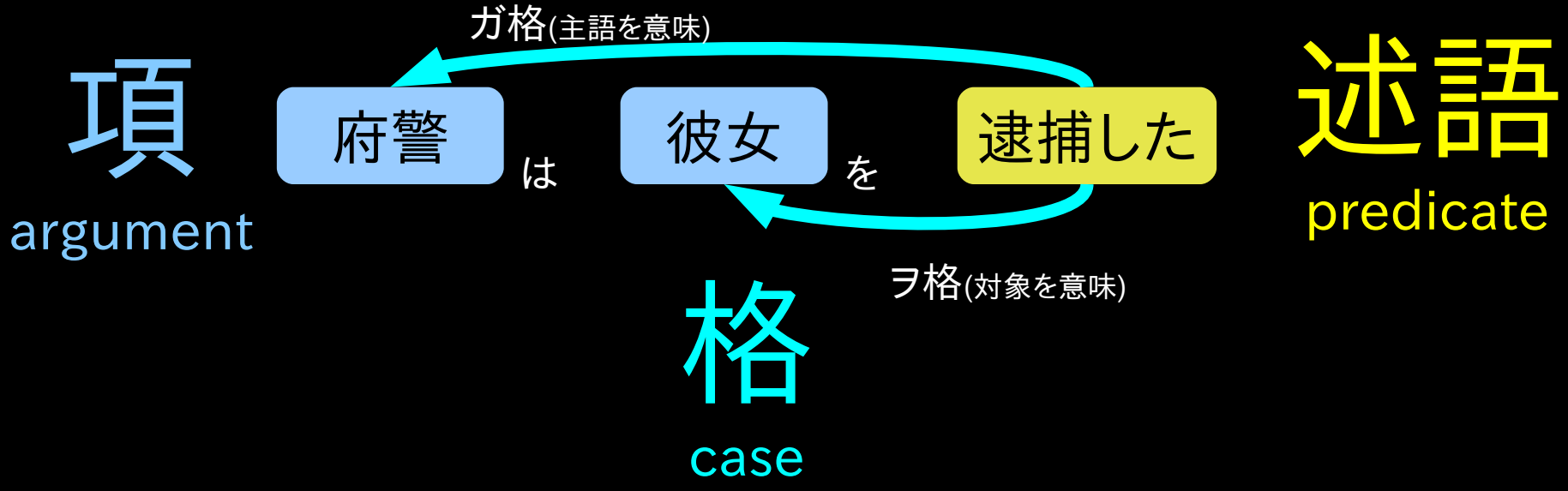
府警は花子を逮捕した.
(○○が)昨日自首したようだ.

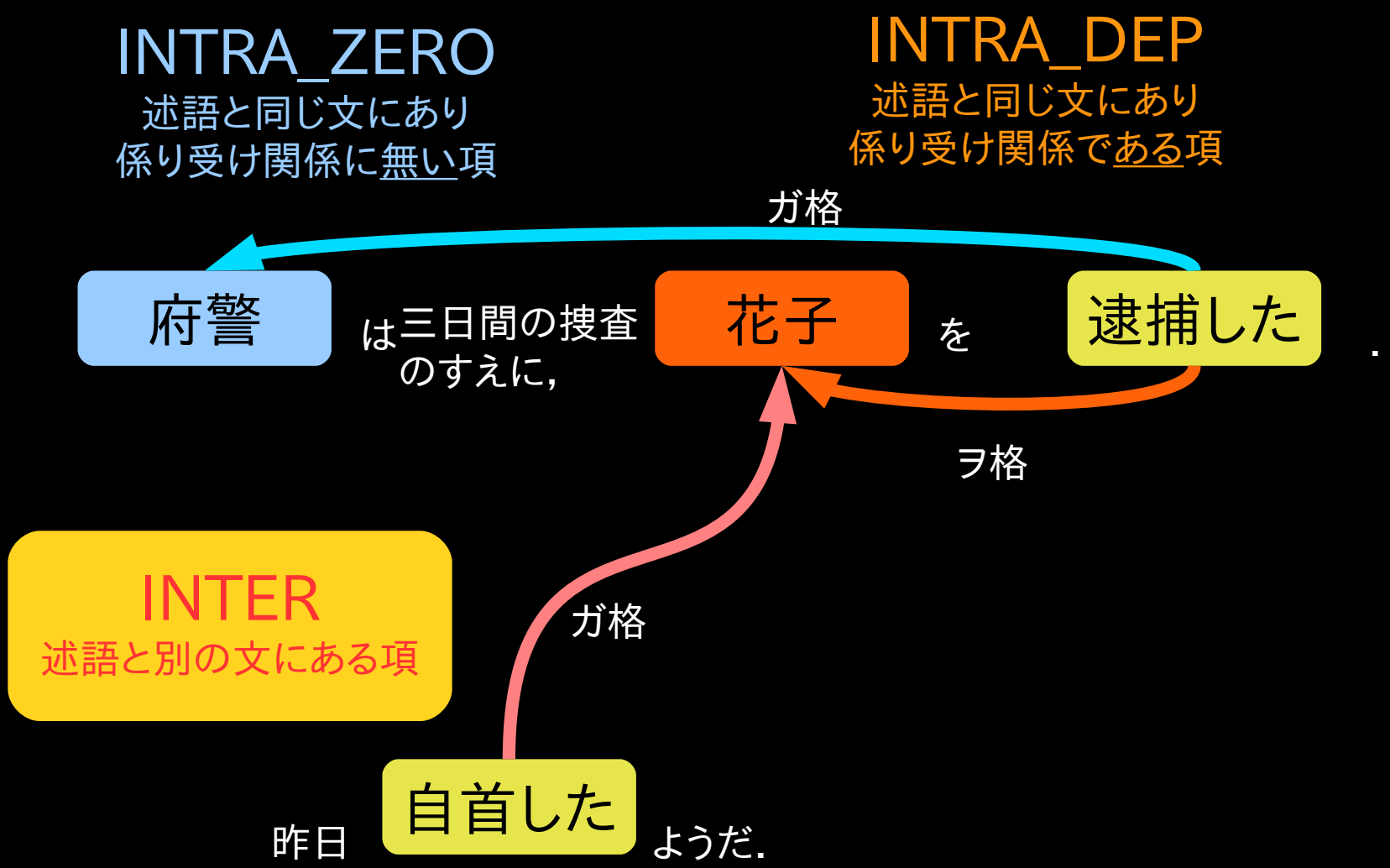
同定する

自動翻訳や情報検索の精度向上に有用

述語項構造解析

Predicate argument structure analysis




INTRA_ZERO
述語と同じ文にあり
係り受け関係に無い項
INTRA_DEP
述語と同じ文にあり
係り受け関係である項
ガ格
府警
は三日間の捜査
のすえに,
花子
を
逮捕した
.
ヲ格
INTER
述語と別の文にある項
ガ格
昨日
自首した
ようだ.

本研究の貢献

府警は彼女を逮捕した.
(○○が)昨日自首したようだ.

(2) それを用いて
省略されたものを見つける

Xを逮捕する
Yが自首する

XとYは
似たものが来やすい

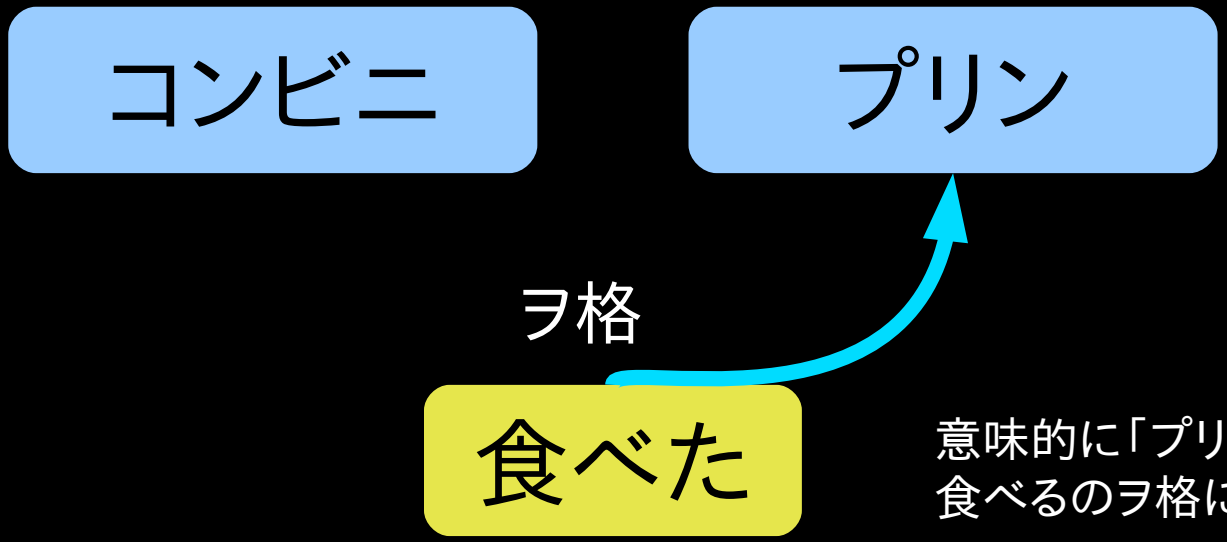
(1)「格構造」の類似度を定義

意味的情報

局所文脈情報

大域文脈情報

コンビニでプリンを買ってきた.
帰ってすぐに(〇〇を)食べた.



意味的に「プリン」の方が
食べるのヲ格になりやすい

1 シソーラス(概念辞書)の利用

名詞	プリン
カテゴリ名	菓子
上位概念	食物

動詞	食べる
カテゴリ名	身体動作
ガ格	人,動物
ヲ格	食物,生物

日本語語彙大系(岩波書店)より

2 コーパス(文書集合)の利用

“プリンを食べる”約1,580,000件
“コンビニを食べる”5件

Googleの検索結果より

意味的情報

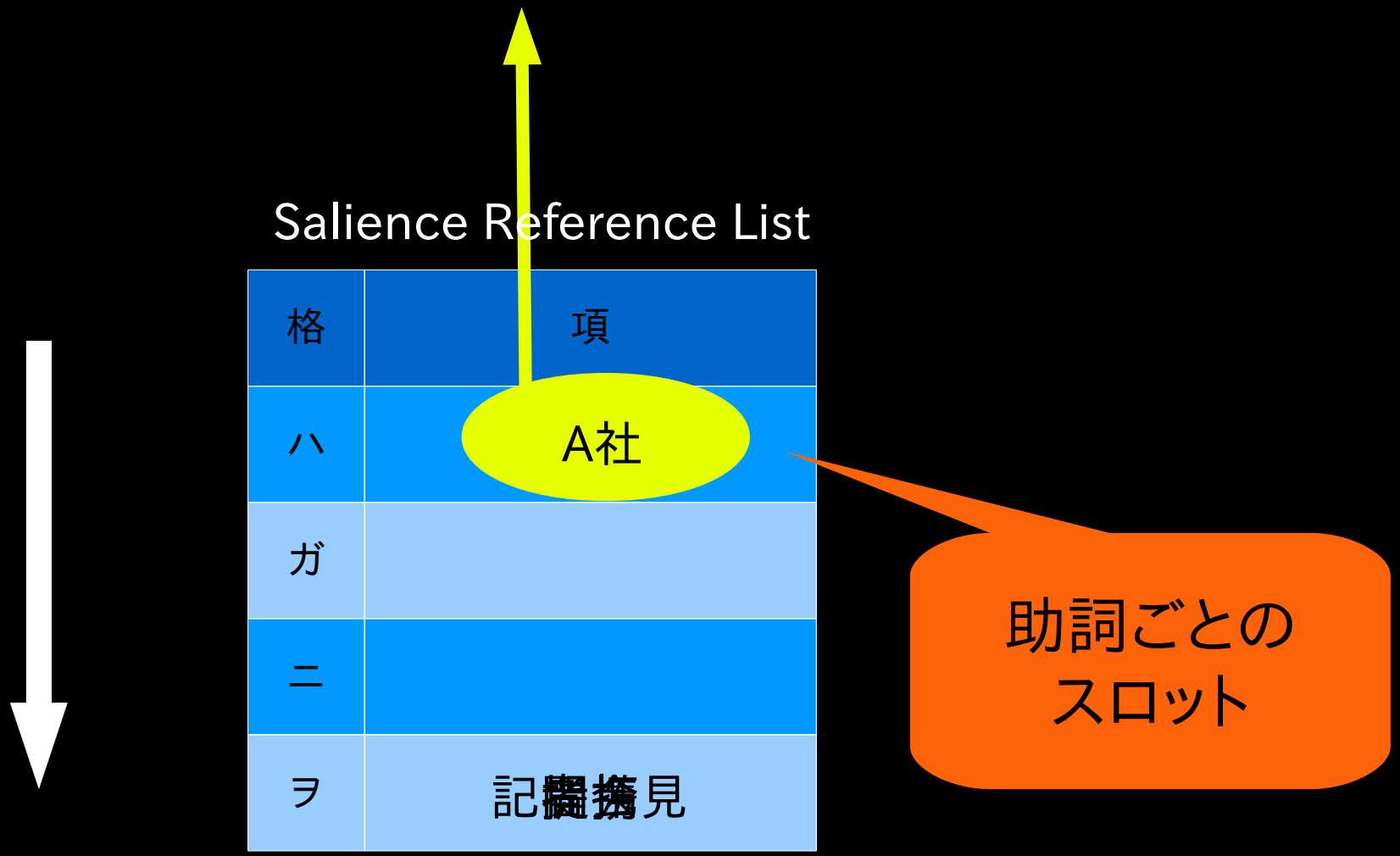
局所文脈情報

大域文脈情報

Salience Reference List

(SRL) [飯田ら03]

A社は記者会見を開き，B社との提携を発表した．
新たな市場の開拓を(〇〇が)目指す．



意味的情報

局所文脈情報

大域文脈情報

項となった回数は文間項同定に有用

A 社は記者会見を開き, B社との提携を発表した.
共同開発により, 新たな市場の開拓を(〇〇が)目指す.

項候補	項になった回数
A社	2　(が-開く, が-発表する)
記者会見	1　（を-開く）
B社	0
提携	1　（を-発表する）

CHAIN_LENGTH素性 [飯田ら03]

項候補が以前の文で何回項になったか

USED素性[Imamura et al.09]

項候補が以前の文で候補になったか否か

意味的情報

局所文脈情報

大域文脈情報

府警は花子を逮捕した.
昨日(〇〇が)自首したようだ.

意味的情報

府警は花子を逮捕した.
昨日(〇〇が)自首したようだ.

“花子が自首する”0件
“府警が自首する”0件

Googleの検索結果より

局所文脈情報

府警は花子を逮捕した.
昨日(〇〇が)自首したようだ.

格	項
ハ	府警
ガ	
ヲ	花子
ニ	

大域文脈情報

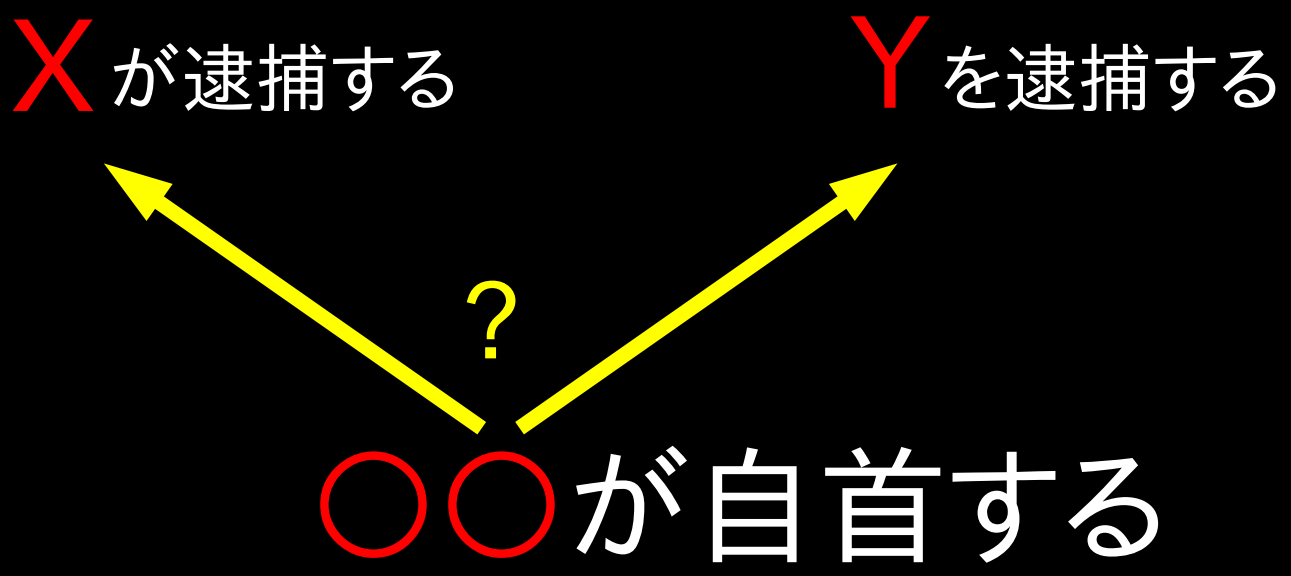
府警は花子を逮捕した.
昨日(〇〇が)自首したようだ.

	使われ方	使用頻度
府警	が逮捕する	1
花子	を逮捕する	1
(その他の候補)	-	0

これらの素性では
判別できない

格構造類似度情報

XがYを逮捕する
(〇〇が)自首する



1599	が自首する
136	者
117	犯人
96	彼
68	人
63	男
36	犯
30	少年
26	高校生
23	梶
22	人間
21	女性
20	息子

5651	が逮捕する
702	署員
698	警察
376	警察官
368	署
230	県警
177	員
153	府警
137	当局
132	警視庁
127	人
126	容疑
115	警官

82112	を逮捕する
15285	人
8484	者
5563	男
2804	名
1188	犯人
1185	男性
763	ら
671	おまえ
587	ところ
562	氏
482	少年
449	人々

webから収集した約5億文[Kawahara et al.09]より作成

格構造類似度

Sim(A,B)≡1-JS(A,B)

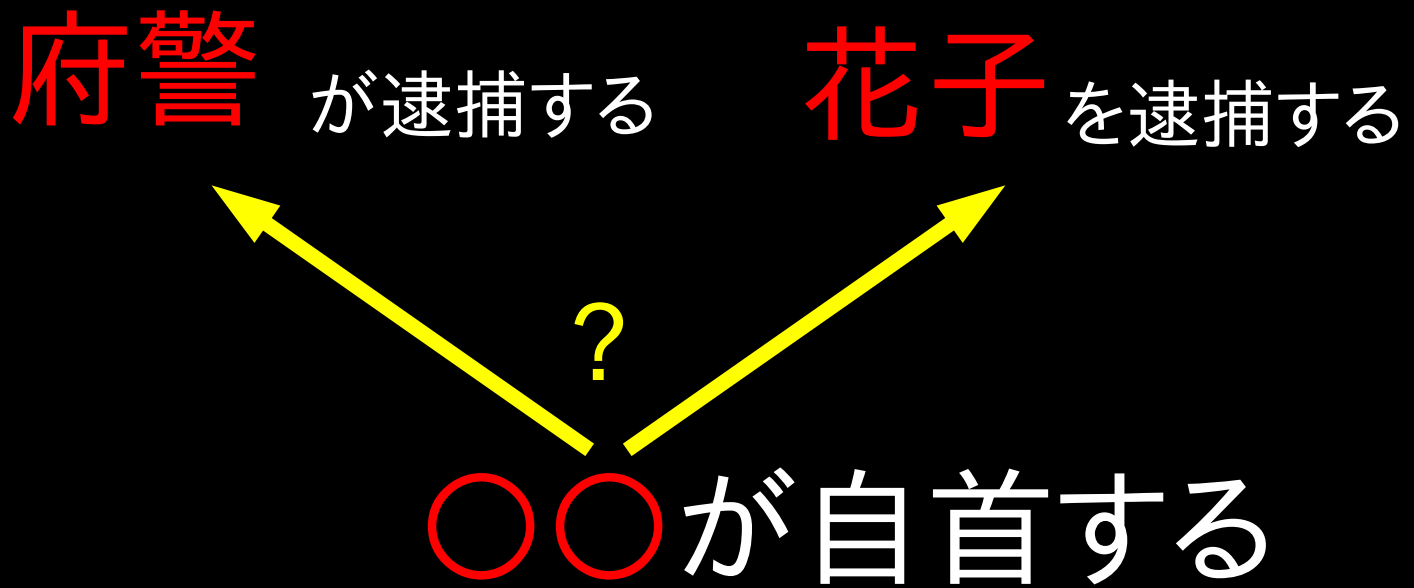
	が自首する	が逮捕する	を逮捕する
が自首する	-	0.4590	0.7568
が逮捕する		-	0.4861
を逮捕する			-

参考

$$JS(p,q) = \frac{1}{2}\left(KL(p,q)+KL(q,p)\right)$$

$$= \frac{1}{2}\left(\sum p(x)\log\frac{p(x)}{\frac{p(x)+q(x)}{2}} + \sum q(x)\log\frac{q(x)}{\frac{p(x)+q(x)}{2}}\right)$$

府警は花子を逮捕した.
昨日(〇〇が)自首したようだ.



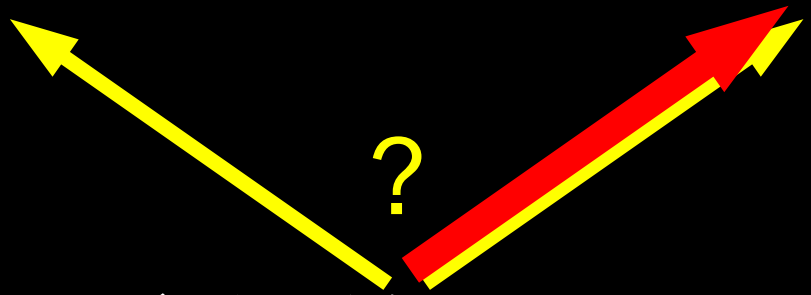
提案手法

警察が花子を逮捕する。
花子が警察に行く。
(○○が)自首する.

先行文脈の情報を用いた解析

警察 が逮捕する 0.2
に行く 0.1

花子 を逮捕する 0.6
が行く 0.1



○○が自首する

(1)項の解析履歴を列挙

(2)“が自首する”との類似度を求める

(3)最大値や平均値を素性に

実験準備

NAISTテキストコーパス

トーナメントモデル

ガ格

就任後初めて地元の大分県へ里帰りしていた村山富市首相は

（中略）

同じ

到着した。

首相は記者団に対し、

（中略）

ガ格

と感想を述べた。

記事

1/1～1/17（約2万文）

社説記事

全て（約2万文）

参考文献[飯田ら10]

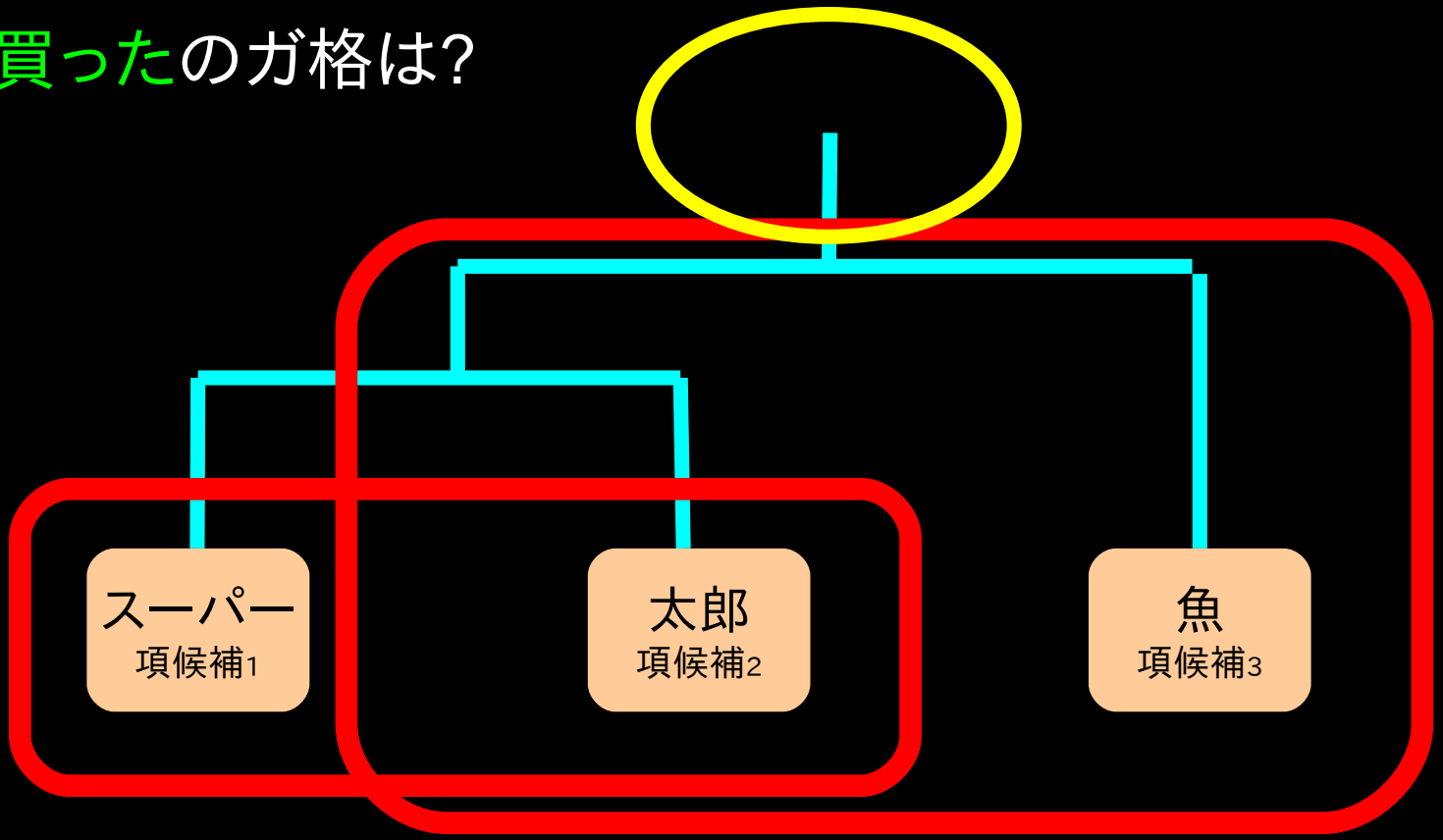
実験準備

NAISTテキストコーパス

トーナメントモデル

スーパーで太郎は魚を買った.

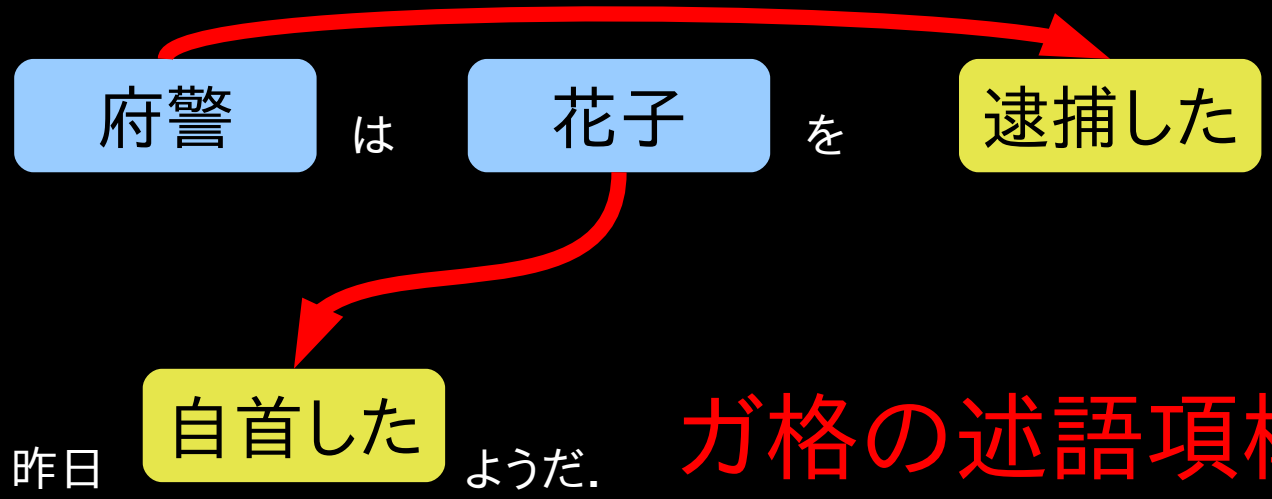
買ったのガ格は?



二値分類
(意味的情報等を用いて分類)

参考文献[飯田ら04]

実験内容



ガ格の述語項構造解析

Baseline
(位置的情報+意味的情報+局所文脈情報)

(A) +CHAIN_LENGTH

(B) +USED

(C) +提案手法

詳細な実験設定

分類器	Support Vector Machine(線形カーネル)
ベースライン素性	[Iida et al.07]の素性
項同定モデル	トーナメントモデル[飯田ら10]
データセット	NAISTテキストコーパス1.4β(2917記事)
テスト方法	5分割交差検定
類似度の学習	Kawahara et al.がwebから収集した約5億文
その他	照応解析と以前の文の述語項構造解析は全て正解したと仮定する

実験結果

提案手法の貢献

コーパスを変えた時の変化

関連手法との相性

A：CHAIN_LENGTH
B：USED
C：提案手法（ウェブテキストより類似度計算）

実験結果

	INTRA_DEP			INTRA_ZERO			INTER		
	P	R	F	P	R	F	P	R	F
Baseline (BL)	78.28	89.63	83.57	43.89	58.38	50.11	17.84	16.59	17.18
BL+A	79.53	91.82	85.23	50.69	60.69	55.24	17.49	21.66	19.34
BL+B	79.53	91.45	85.07	51.80	59.85	55.53	18.17	24.76	20.95
BL+C	79.87	92.33	85.65	60.07	63.09	61.54	17.30	27.85	21.33

（1）関連手法よりも精度向上の幅が大きい

（2）INTERのみならずINTRA_ZEROにも提案手法は有効

※予稿の数値とは異なります(修正しています)

実験結果

提案手法の貢献

コーパスを変えた時の変化

関連手法との相性

A : CHAIN_LENGTH
B : USED
C : 提案手法（ウェブテキストより類似度計算）
D : 提案手法（毎日新聞10年より類似度計算）

実験結果

	INTRA_D			INTRA_Z			INTER		
	P	R	F	P	R	F	P	R	F
Baseline	78.28	89.63	83.57	43.89	58.38	50.11	17.84	16.59	17.18
BL+A	79.53	91.82	85.23	50.69	60.69	55.24	17.49	21.66	19.34
BL+B	79.87	92.33	85.65	60.07	63.09	61.54	17.30	27.85	21.33
BL+C	85.34	93.47	89.22	57.98	63.48	60.60	16.55	28.06	20.81
BL+D	85.55	93.81	89.49	61.58	64.28	62.90	17.07	31.10	22.04
BL+C+D									

(1)Dは特にINTRA_Dの精度向上に貢献

(2)CとDは相補的にはたらく

実験結果

提案手法の貢献

コーパスを変えた時の変化

関連手法との相性

A : CHAIN_LENGTH
B : USED
C : 提案手法（ウェブテキストより類似度計算）
D : 提案手法（毎日新聞10年より類似度計算）

実験結果

	INTRA_DEP			INTRA_ZERO			INTER		
	P	R	F	P	R	F	P	R	F
BL	78.28	89.63	83.57	43.89	58.38	50.11	17.84	16.59	17.18
BL+A+B	79.54	91.49	85.10	51.62	59.78	55.40	18.40	24.92	21.16
BL+A+B+C	79.54	92.31	85.45	59.77	62.34	61.03	18.18	29.38	22.46
BL+A+B+C+D	85.42	93.49	89.27	61.00	63.20	62.08	16.91	31.27	21.94

提案手法は関連手法とも相補的にはたらく

エラー分析

述語の曖昧性

機能動詞結合

述語の曖昧性

数字を早急に詰める必要性を強調した。

ジャムをビンに詰める。

複数の意味を持つ述語がある

「～を詰める」のパターン

もの
雪
荷物
物
品
香り
ミ
根
封筒
葉
商品
綿

食材
ごはん
酒
クリーム
水
豆
野菜
具
原酒

間合い
距離
差
息
内容
部分
材
等
間

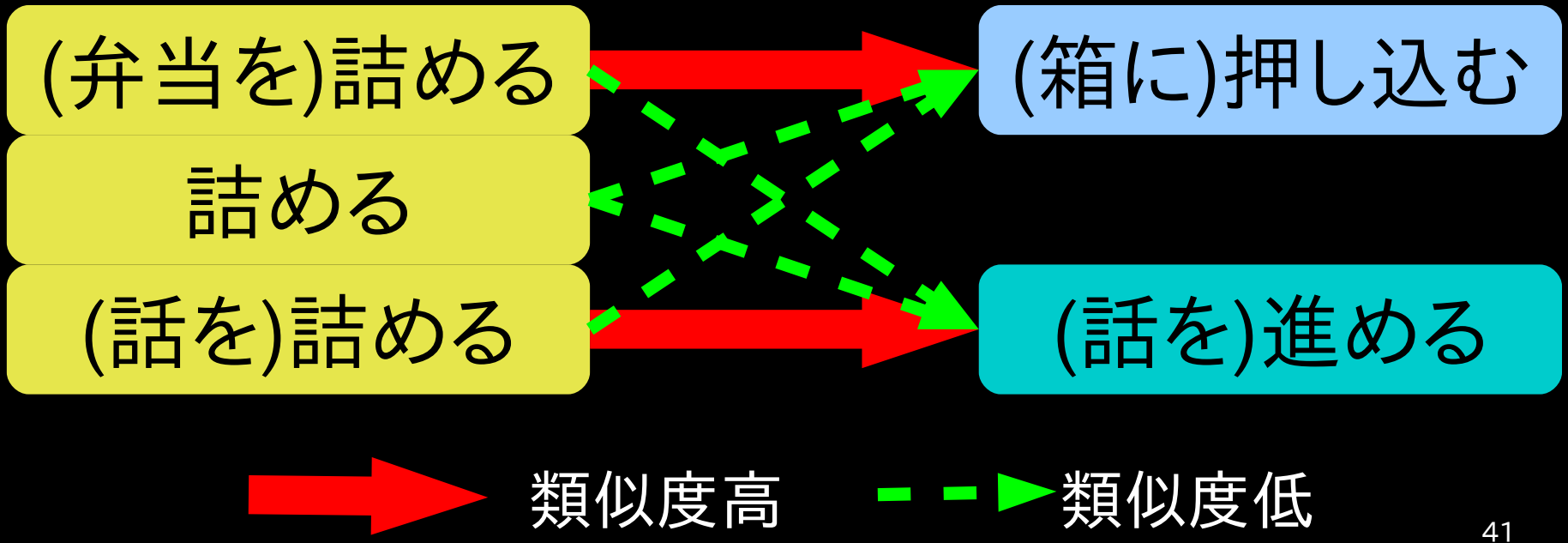
など
%
だけ
」
こと
ところ

webから収集した大量の述語対(約11億対)より
頻度上位のものから作成

多義語の場合類似度がうまくとれない

(1)容器に物をすき間がないように入れる
(6)究極まで進める。

三省堂 大辞林より



エラー分析

述語の曖昧性

機能動詞結合

機能動詞結合

名詞+格助詞+機能動詞

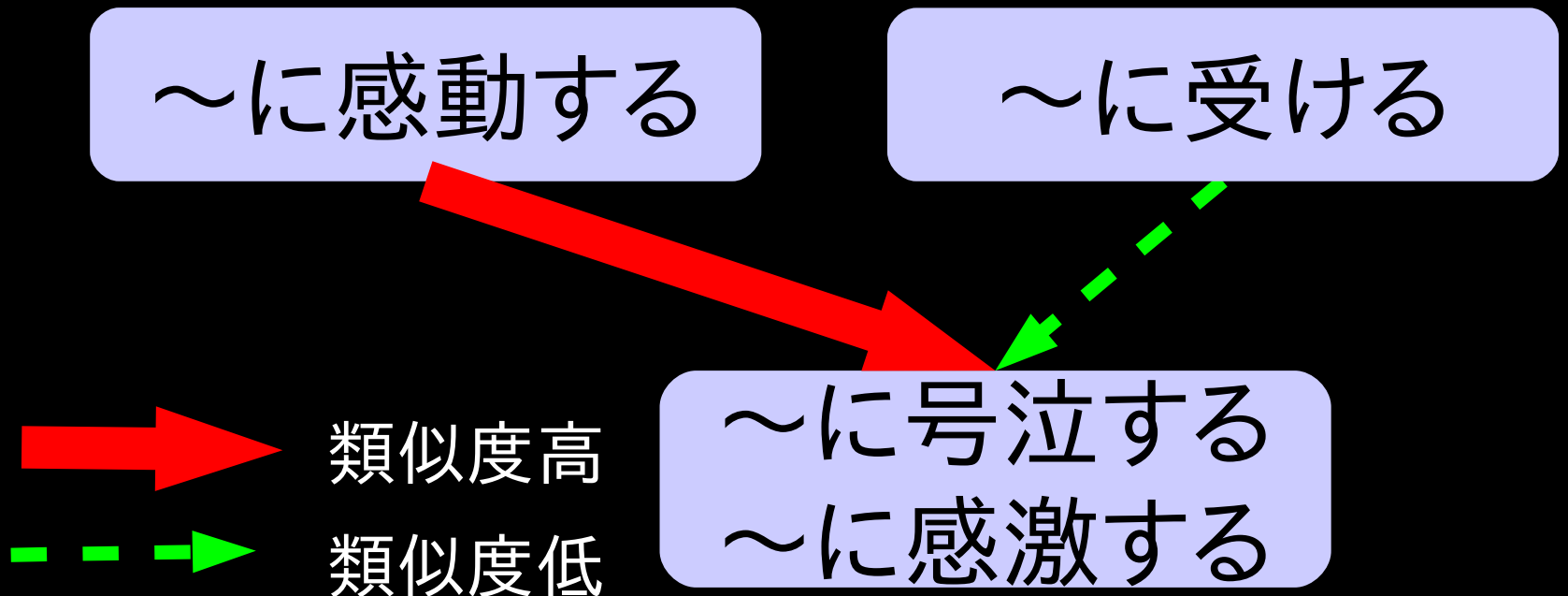
情熱的な演技に感動を受ける

主たる意味は名詞にある

動詞の自立性は希薄

機能動詞結合
⇒類似度がうまくとれない

情熱的な演技に感動を受ける



まとめ

「格構造類似度」を定義

述語項構造解析に有効

今後の課題

類似度では捉えられない「文脈」を捉える