

格フレーム辞書の 漸次的自動構築

[自然言語処理, Vol.12, No.2, pp.109-131, 2005]

河原 大輔 黒橋 禎夫
京都大学 大学院情報学研究科

自然言語理解 ↔ 知識



クロールで泳いでいる女の子を見た
望遠鏡で泳いでいる女の子を見た



格フレーム

{人,子,...}が
{クロール,平泳ぎ,...}で
{海,大海,...}を泳ぐ

{人,者,...}が
{双眼鏡,望遠鏡,...}で
{姿,人,...}を見る

→ 格フレーム辞書を自動構築

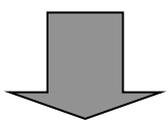
複雑な言語現象

- 係助詞句
その車は速い
あの本も読んだ
- 連体修飾
速い車
読んだ本
- 二重主語構文
その車はエンジンがよい
- 外の関係
魚を焼くけむり
- 格変化
社会党が新進党の支持を得る
社会党が新進党から支持を得る

格フレーム自動構築のアプローチ

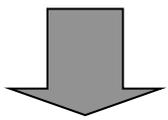
- 大量の生テキストから格フレームを自動構築
- 自動解析結果に含まれる構文的・意味的曖昧性に対処
 - 構文的曖昧性 (構文解析誤り)
 - 信頼度の高い部分のみを用いる
 - 意味的曖昧性
 - クラスタリング
- 漸次的格フレーム構築
 - 様々な関係を段階的に得る(ブートストラップ)

2600万文
(新聞記事26年分)



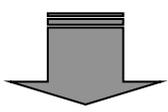
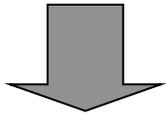
構文解析

精度:90%

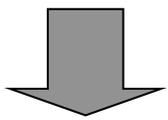


フィルタリング

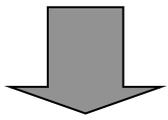
精度:98% for
全係り受けの20%



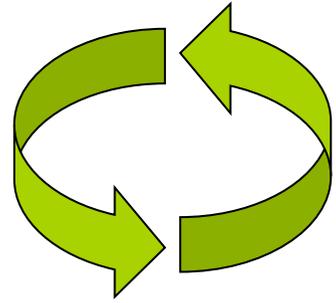
述語項構造



クラスタリング

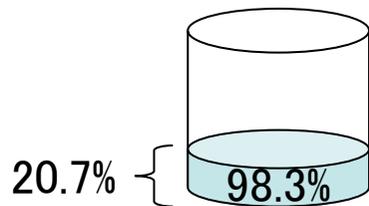


格フレーム
(2万述語)



問題1: 構文的曖昧性

- 信頼度の高い部分のみを用いる
 - ひどい 風邪を 引いたので、家で 寝ていた。
 - 被害者を 早く 救い出すべきだ。
 - 火の 回りが 早く 救いだせなかった。
 - その 議員は 法案を 提出した。
 - 議員が 提出している 法案は ...



係り受けの20.7%について
98.3%の精度で抽出

問題2: 意味的曖昧性

従業員 が 車 に 荷物 を 積む

作業員 が 荷物 を 積む

飛行機 に 荷物 を 積む

彼 が 車 に 物資 を 積む

トラック に 物資 を 積む

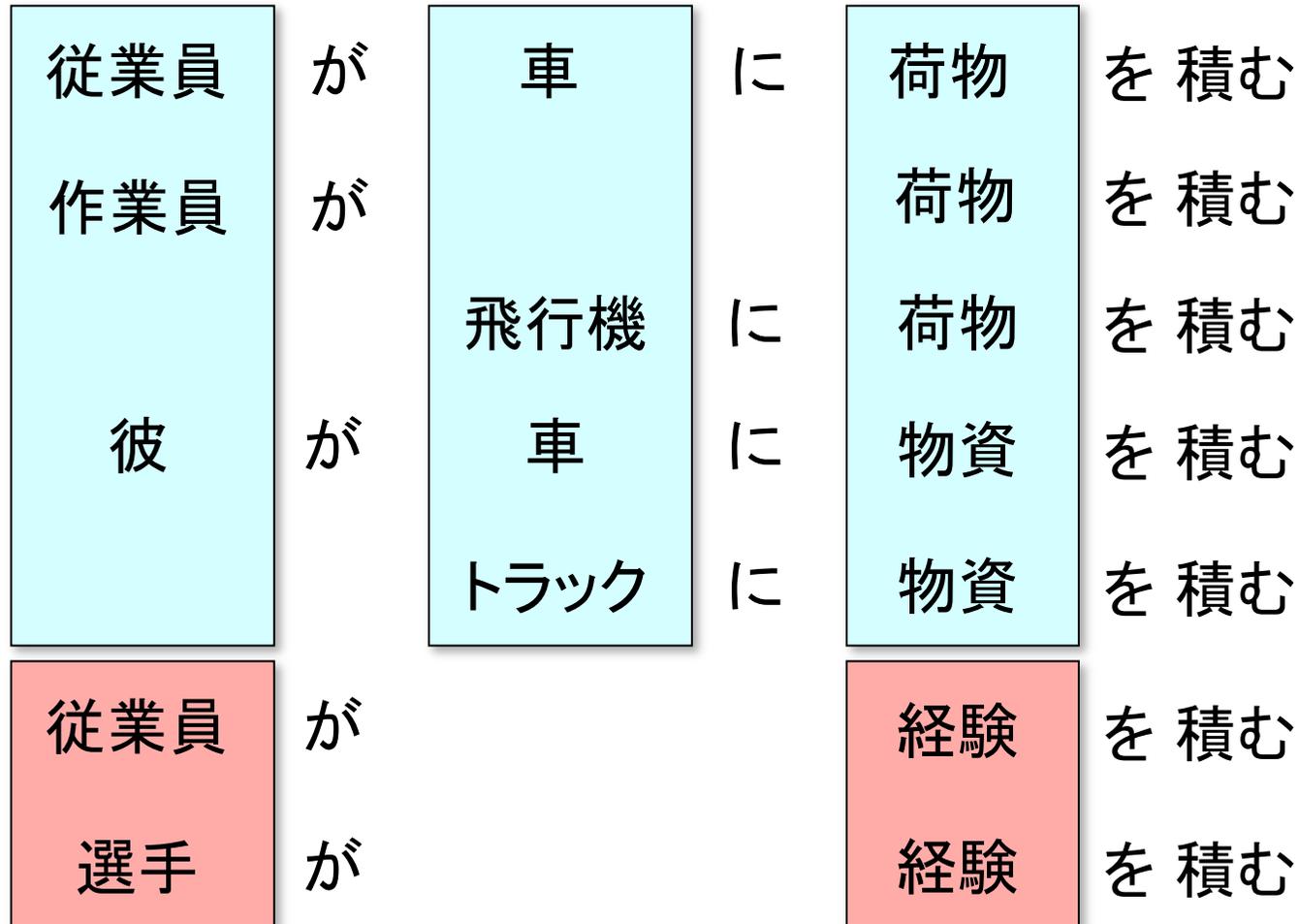
従業員 が 経験 を 積む

選手 が 経験 を 積む

直前格ごとにまとめる

従業員	が	車	に	荷物	を積む
作業員	が			荷物	を積む
		飛行機	に	荷物	を積む
彼	が	車	に	物資	を積む
		トラック	に	物資	を積む
従業員	が			経験	を積む
選手	が			経験	を積む

クラスタリング



格フレームの漸次的自動構築

- この法案はA議員が提出した。

{議員, 委員...}が {法律, 案...}を 提出する

- その車はエンジンがよい。 ⇒ 二重主語構文

{エンジン}が よい

- 法案を提出した議員...

{議員, 委員...}が {法律, 案...}を 提出する

- 法案を提出する見通し... ⇒ 外の関係

{議員, 委員...}が {法律, 案...}を 提出する

格フレーム検索

経験 検索 ヘルプ

名詞から検索, CSV出力(名詞検索のみ), 上位 個までの単語を表示, 頻度 以上の単語を表示

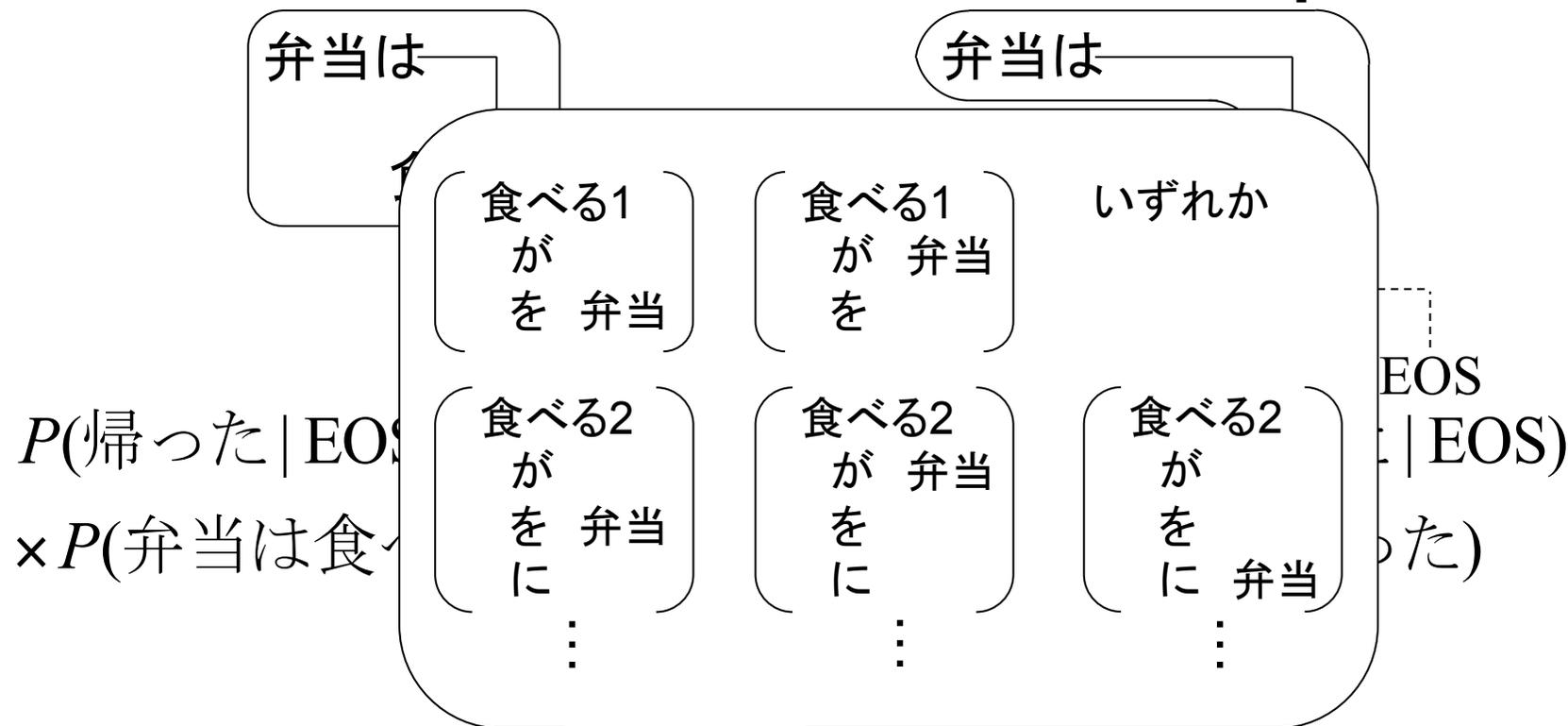
検索例: 「積む」, 「修行を 積む」, 「積む:P」(受身), 「積む:C」(使役)

「経験/けいけん」を検索しました。

格フレームID	格	頻度
有る/ある:動25	ガ格	70992
生かす/いかす:動2	ヲ格	45320
積む/つむ:動1	ヲ格	37342
する/する:動56	ヲ格	34467
持つ/もつ:動20	ヲ格	28286
無い/ない:形22	ガ格	18996
有する/ゆうする:動4	ヲ格	15417
出来る/できる:動41	ガ格	7979
踏まえる/ふまえる:動8	ヲ格	6520
必要だ/ひつようだ:形39	ガ格	5835
有す/ゆうす?有する/ゆうする:動6	ヲ格	5232
重ねる/かさねる:動8	ヲ格	4610
する/する:動:P31	ヲ格	3788
経る/へる:動19	ヲ格	3384
する/する:動:C14	ヲ格	2940
方/ほう:判2	ヲ格	2724
必要だ/ひつようだ+なる/なる:動16	ガ格	2665
豊富だ/ほうふだ:形6	ガ格	2663

構文・格解析の統合的確率モデル

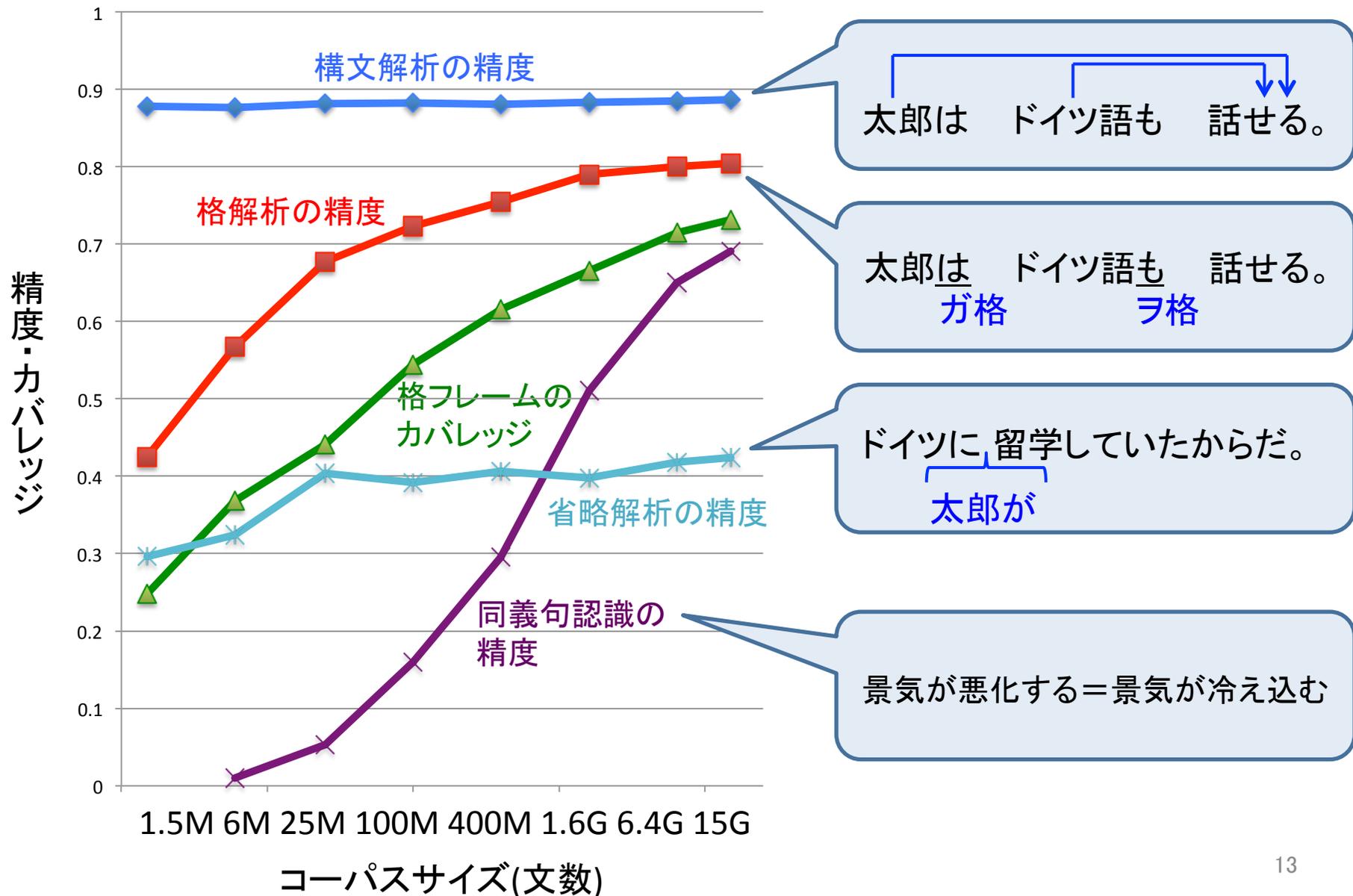
[河原・黒橋, 2007]



$$P(C_i | b_{i-1}, f_{i-1}, w_{h_i}, f_{h_i}) \approx P(CS_i | f_i, w_{h_i}) \times P(f_i | f_{h_i})$$

述語項構造生成確率 用言タイプ生成確率

超大規模コーパスによる言語解析の高度化



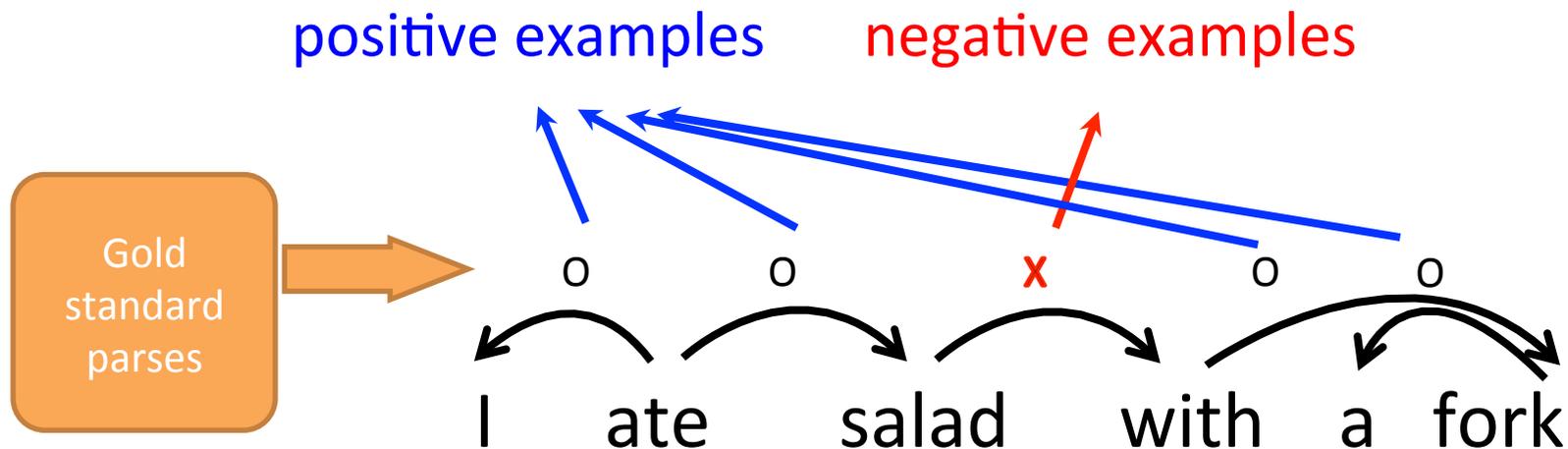
格フレーム構築の 一般化と他言語への展開

- 構文的曖昧性
 - フィルタリングルール → モデル化 (学習)
- 意味的曖昧性
 - 類似度に基づく凝集型クラスタリング → non-parametric ベイズモデル
 - 拡張: 格フレームに加えて意味クラスを学習

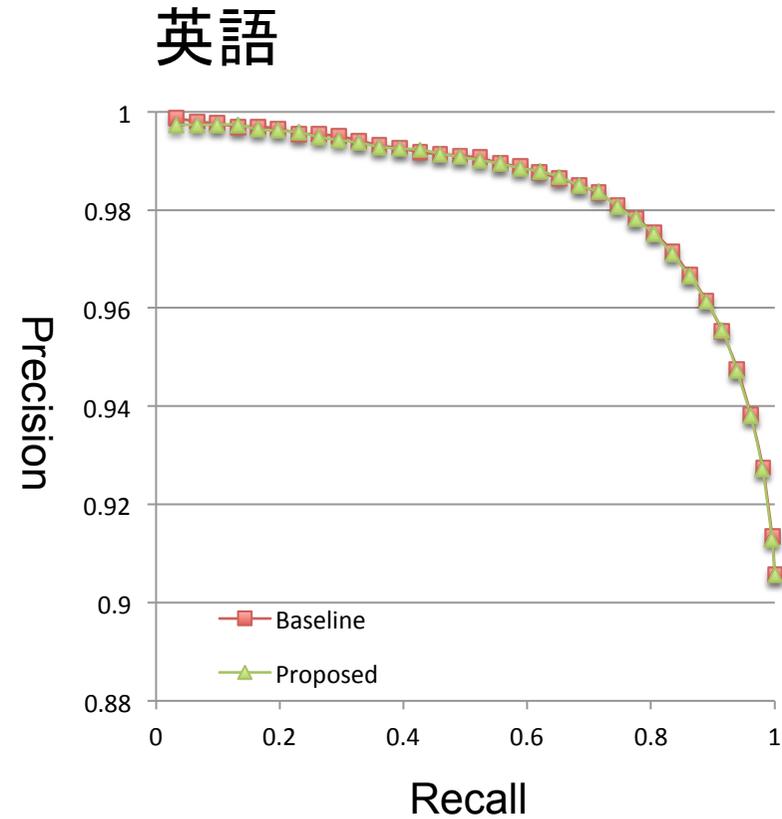
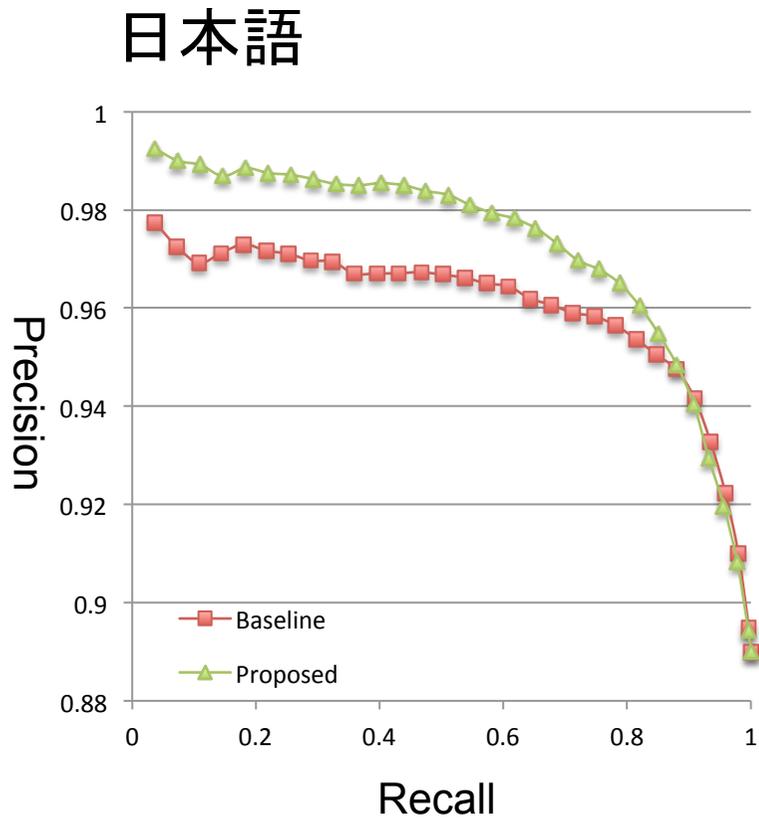
確信度の高い係り受けの自動抽出

[金ら, 自然言語処理, 2014 (to appear)]

- 自動解析で得られた係り受けの確からしさを推定
- 係り受け解析の学習で用いるタグ付きコーパス (treebank)を利用して学習

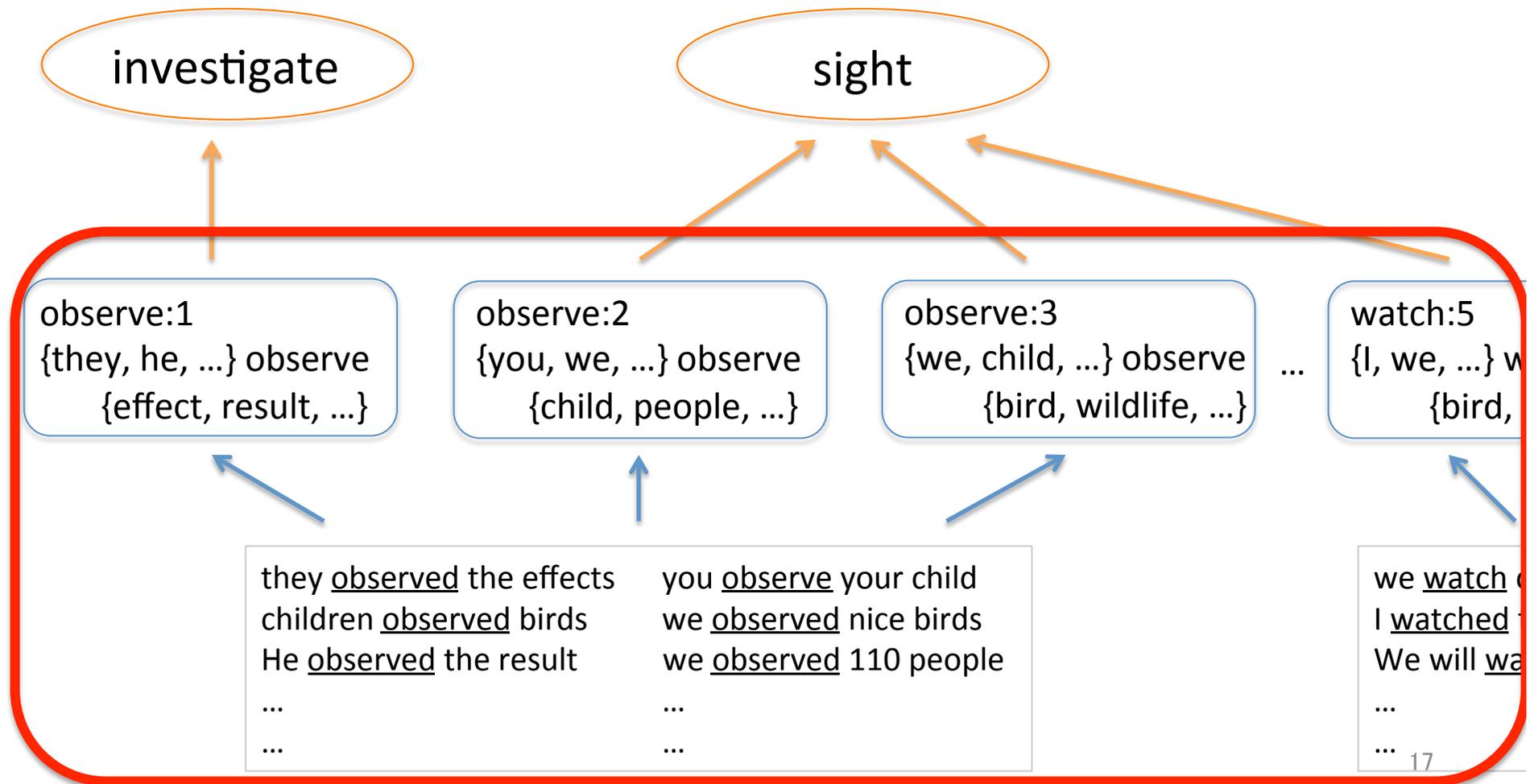


日本語・英語における 係り受け抽出の実験結果



格フレームと意味クラスの自動獲得

[Kawahara et al., ACL2014]



- The doctor observed the **effects** of ...
- This statistical ability to observe an **effect** ...
- They did not observe a residual **effect** of ...
- We could observe the **results** at the same time ...
- My first opportunity to observe the **results** of ...
- You can observe beautiful **birds** ...
- children may then observe **birds** ...
- ...

nsubj:{they, doctor, ..} observe dobj:{effect}
nsubj.they:825 dobj.effect:5,070
nsubj.doctor:235
prep_at:{time, point, ..}
prep_at.time:71
prep_at.point:20

nsubj:{doctor, we, ..} observe dobj:{result}
nsubj.doctor:531 dobj.result:2,320
nsubj.we:291
prep_at:{time, point, ..}
prep_at.time:93
prep_at.point:37

nsubj:{you, child, ..} observe dobj:{bird}
nsubj.you:704 dobj.bird:1,692
nsubj.child:563

Chinese Restaurant Process

$$P(c_j | f_i) \propto \begin{cases} \frac{n(c_j)}{N + \alpha} \cdot P(f_i | c_j) & j \neq new \\ \frac{\alpha}{N + \alpha} \cdot P(f_i | c_j) & j = new \end{cases}$$

$$P(f_i | c_j) = \prod_{w \in W} P(w | c_j)^{\text{count}(f_i, w)}$$

f_i : initial frame

c_j : cluster
(semantic frame)

$$P(w | c_j) = \frac{\text{count}(c_j, w) + \beta}{\sum_{v \in W} \text{count}(c_j, v) + |W| \cdot \beta}$$

nsubj:{they, doctor, ..} observe dobj:{effect, result}
nsubj.they:825 dobj.effect:5,070
nsubj.doctor:766 dobj.result:2,320
nsubj.we:291
prep_at:{time, point, ..}
prep_at.time:164
prep_at.point:57

nsubj:{you, child, ..} observe dobj:{bird}
nsubj.you:704 dobj.bird:1,692
nsubj.child:563

Case frame search

Case frame search

orchid.kuee.kyoto-u.ac.jp/~kawahara/cf-search/english.crp.gigaword/

Case frame search

Verb: search reset

Number of displayed instances:

"observe" has 16 frames.

ID: observe:1

nsubj <name>:7296, he:1656, she:449, we:253, they:234, i:187, it:179, report:121, official:114, analyst:110, president:73, ...
 ccomp <s>:14587
 prep_in <name>:100, interview:69, book:32, report:23, speech:19, article:16, 1960s:13, editorial:12, statement:11, commentary:10, scenario:10, ...

ID: observe:2

nsubj <name>:1227, people:182, they:162, he:156, country:135, muslim:131, nation:125, we:99, state:77, it:76, most:69, ...
 dobj <name>:1221, day:932, anniversary:865, holiday:453, month:227, time:188, birthday:144, ceremony:129, ban:128, festival:102, fast:95, ...
 prep_on <name>:290, day:15, subway:6, spot:6, treatment:6, complex:4, display:4, sale:3, control:3, protection:2, test:2, ...
 prep_in <name>:132, way:12, city:10, honour:9, memory:9, 1675:7, state:6, protest:5, town:4, day:4, month:4, ...

ID: observe:3

nsubj <name>:1035, group:402, rebel:243, it:240, they:235, side:228, government:196, force:145, troop:114, faction:110, militants:79, ...
 dobj cease-fire:2789, truce:1461, period:355, agreement:283, cessation:48, <name>:47, deal:43, gorilla:40, which:36, suspension:32, border:27, ...
 prep_since <name>:372, 1997:71, 1994:8, end:6, beginning:3, start:3, war:2, eve:2
 prep_in <name>:85, support:38, campaign:12, region:11, territory:9, fight:8, theory:8, war:7, line:6, area:5, republic:5, ...

※ English Gigaword Corpus (40億語) から獲得

まとめ

- 自然言語理解において重要な知識源となる
格フレームを生コーパスから自動構築した
 - 漸次的なアプローチ
 - 超大規模コーパスの利用
- 構文・格解析、省略解析、機械翻訳などの精度向上に寄与している



<http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学格フレーム>