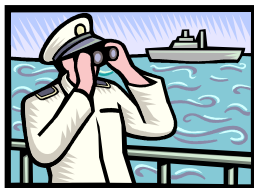
格フレーム辞書の漸次的自動構築

[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

言語処理学会20周年記念シンポジウム (2014/10/4)

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**

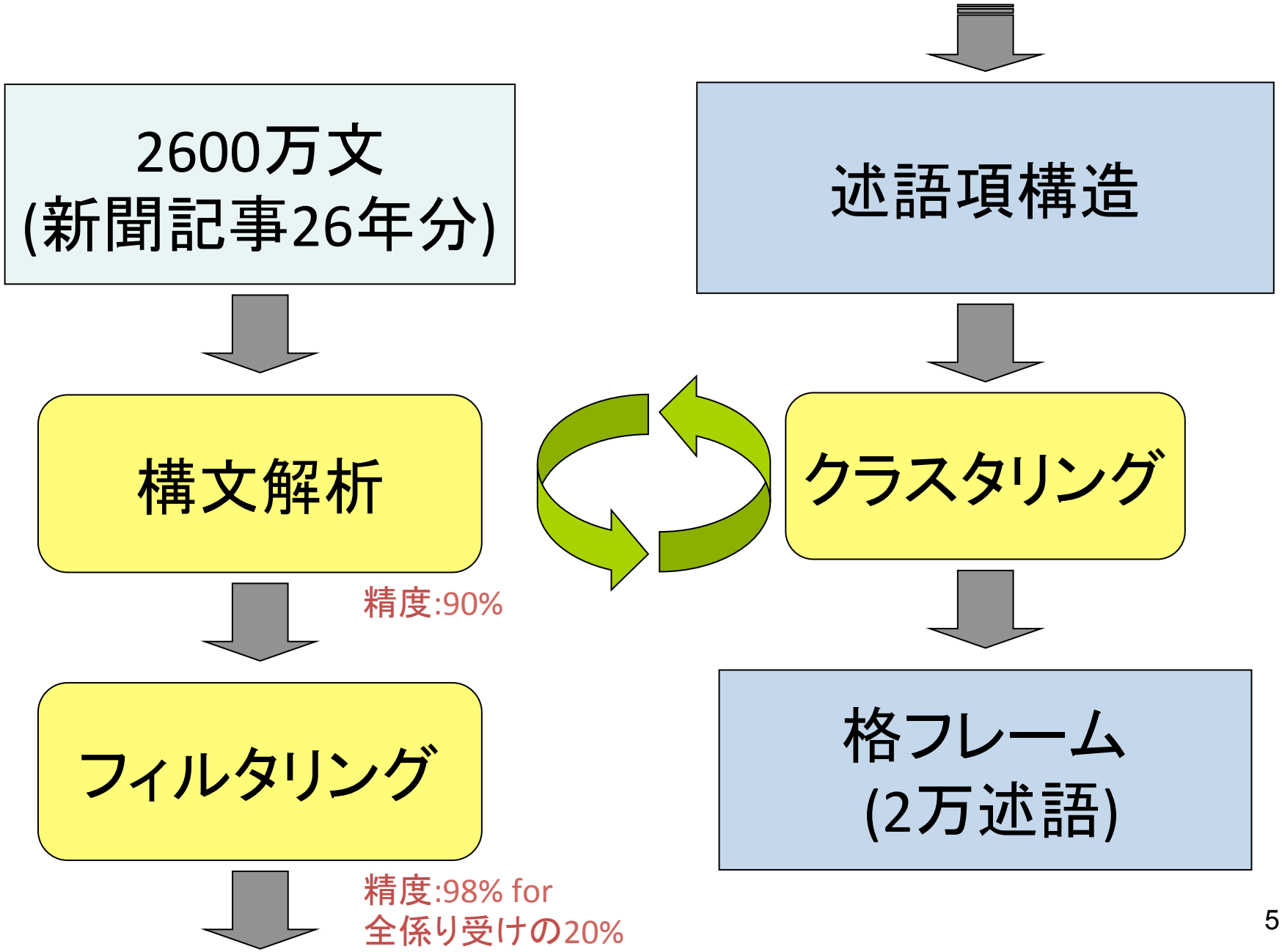
→ 格フレーム辞書を自動構築

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

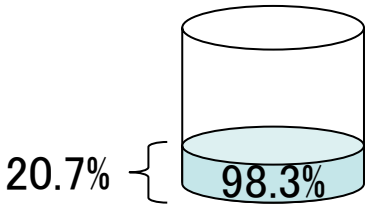
格フレーム自動構築のアプローチ

- 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

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

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

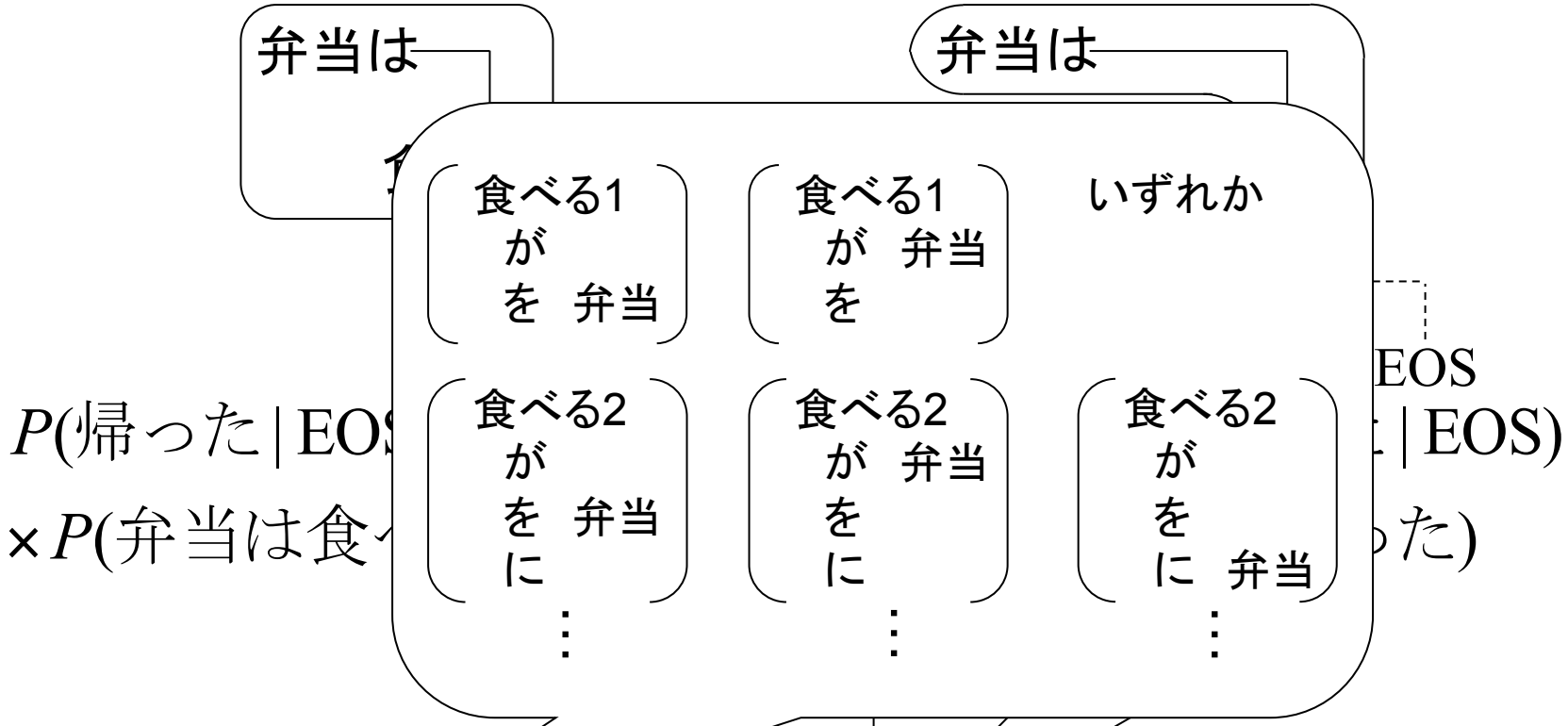
- 法案を提出する見通し．．． ⇒ 外の関係

{議員，委員．．}が　{法律，案．．}を　提出する

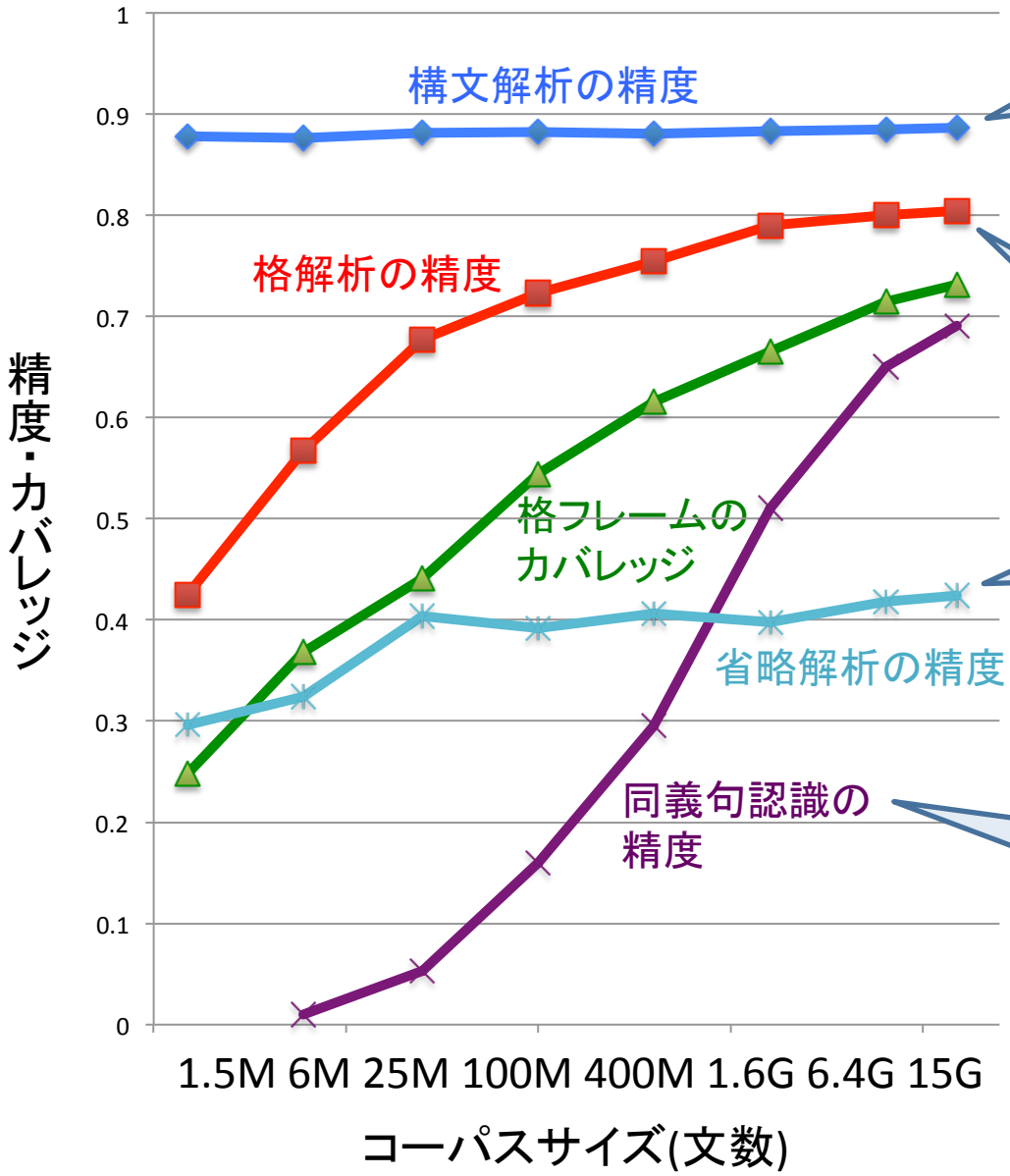
格フレームID	格	頻度
有る/ある:動25	ガ格	70992
生かす/いかす:動2	ヲ格	45320
積む/つむ:動1	ヲ格	37342
する/する:動56	ヲ格	34467
持つ/もつ:動20	ヲ格	28286
無い/ない:形22	ガ格	18996
有する/ゆうする:動4	ヲ格	15417
出来る/できる:動41	ガ格	7979
踏まえる/ふまえる:動8	ヲ格	6520
必要だ/ひつようだ:形39	ガ格	5835
有す/ゆうす?有する/ゆうする:動6	ヲ格	5232
重ねる/かさねる:動8	ヲ格	4610
する/する:動:P31	ヲ格	3788
経る/へる:動19	ヲ格	3384
する/す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



超大規模コーパスによる言語解析の高度化



太郎は　ドイツ語も　話せる。

太郎は　ドイツ語も　話せる。
ガ格　ヲ格

ドイツに 留学していたからだ。
太郎が

景気が悪化する＝景気が冷え込む

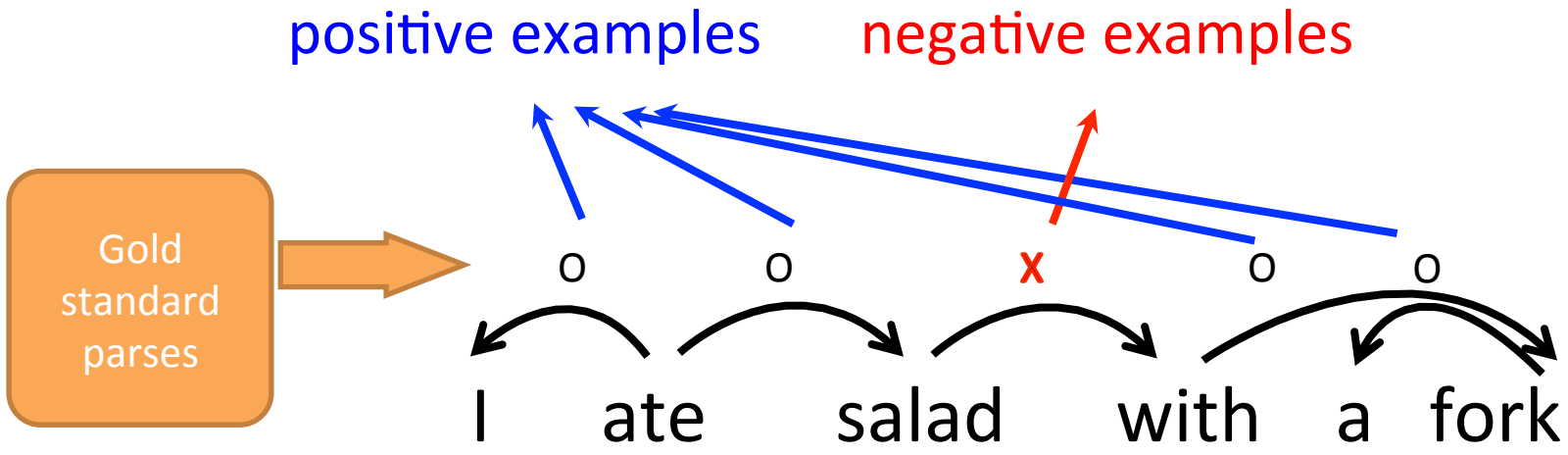
格フレーム構築の一般化と他言語への展開

- 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

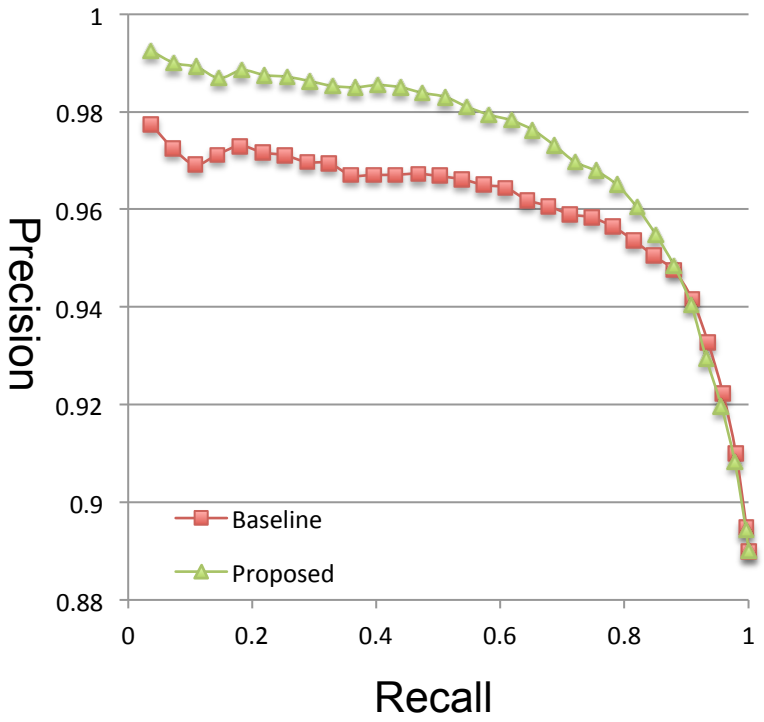
[金ら, 自然言語処理, 2014 (to appear)]

- 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

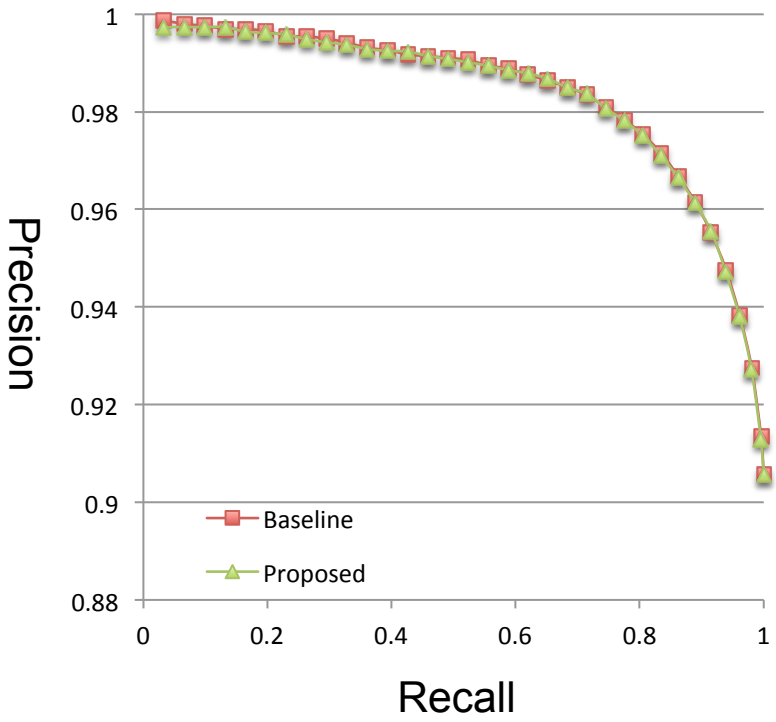


日本語・英語における係り受け抽出の実験結果

日本語

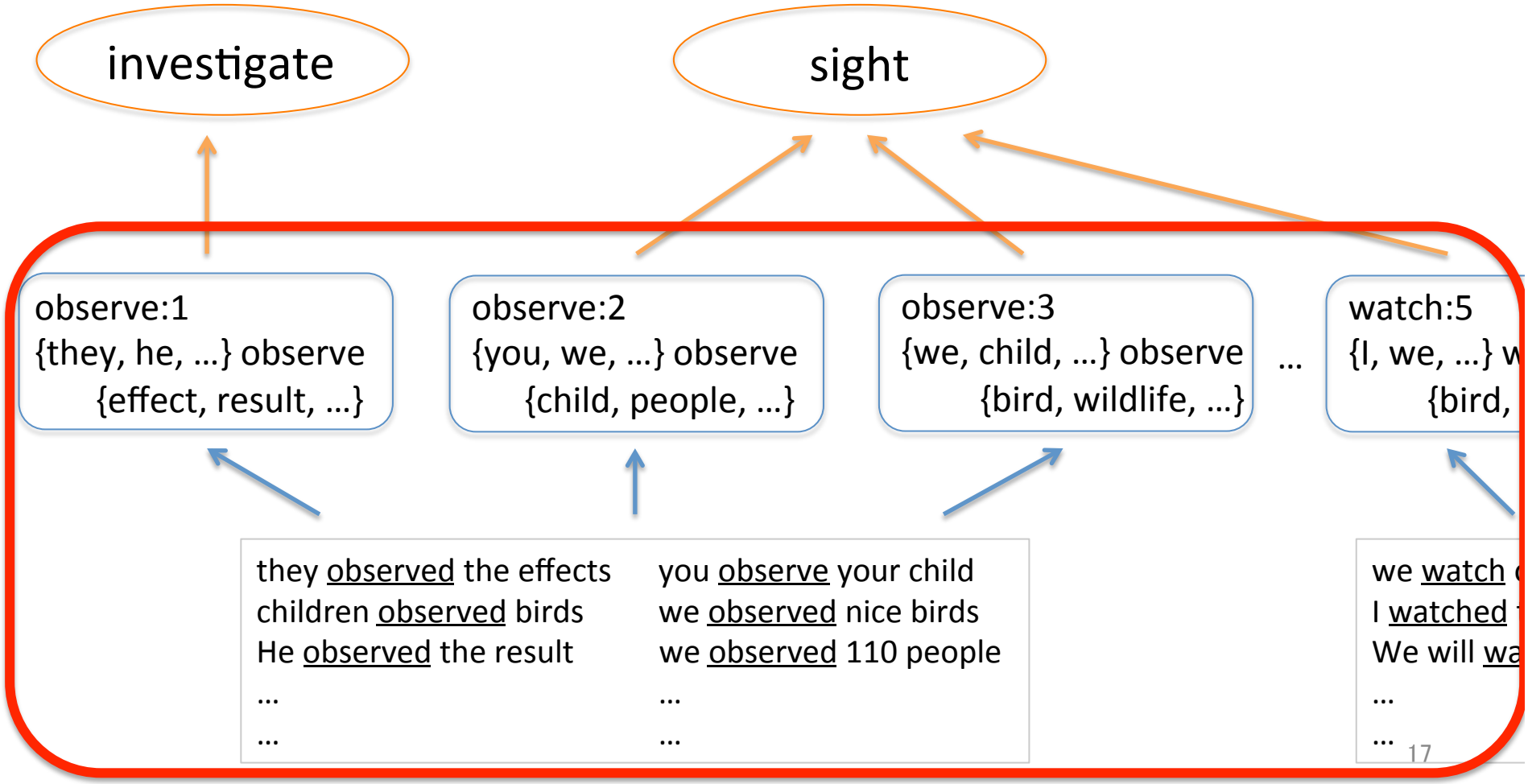


英語



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

nsubj.doctor:235

dobj.effect:5,070

prep_at:{time, point, ..}

prep_at.time:71

prep_at.point:20

nsubj:{doctor, we, ..} observe dobj:{result}

nsubj.doctor:531

nsubj.we:291

dobj.result:2,320

prep_at:{time, point, ..}

prep_at.time:93

prep_at.point:37

nsubj:{you, child, ..} observe dobj:{bird}

nsubj.you:704

nsubj.child:563

dobj.bird:1,692

Chinese Restaurant Process

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

f_i : initial frame

c_j : cluster (semantic frame)

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

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

nsubj:{they, doctor, ..} observe dobj:{effect, result}

nsubj.they:825
nsubj.doctor:766
nsubj.we:291

dobj.effect:5,070
dobj.result:2,320

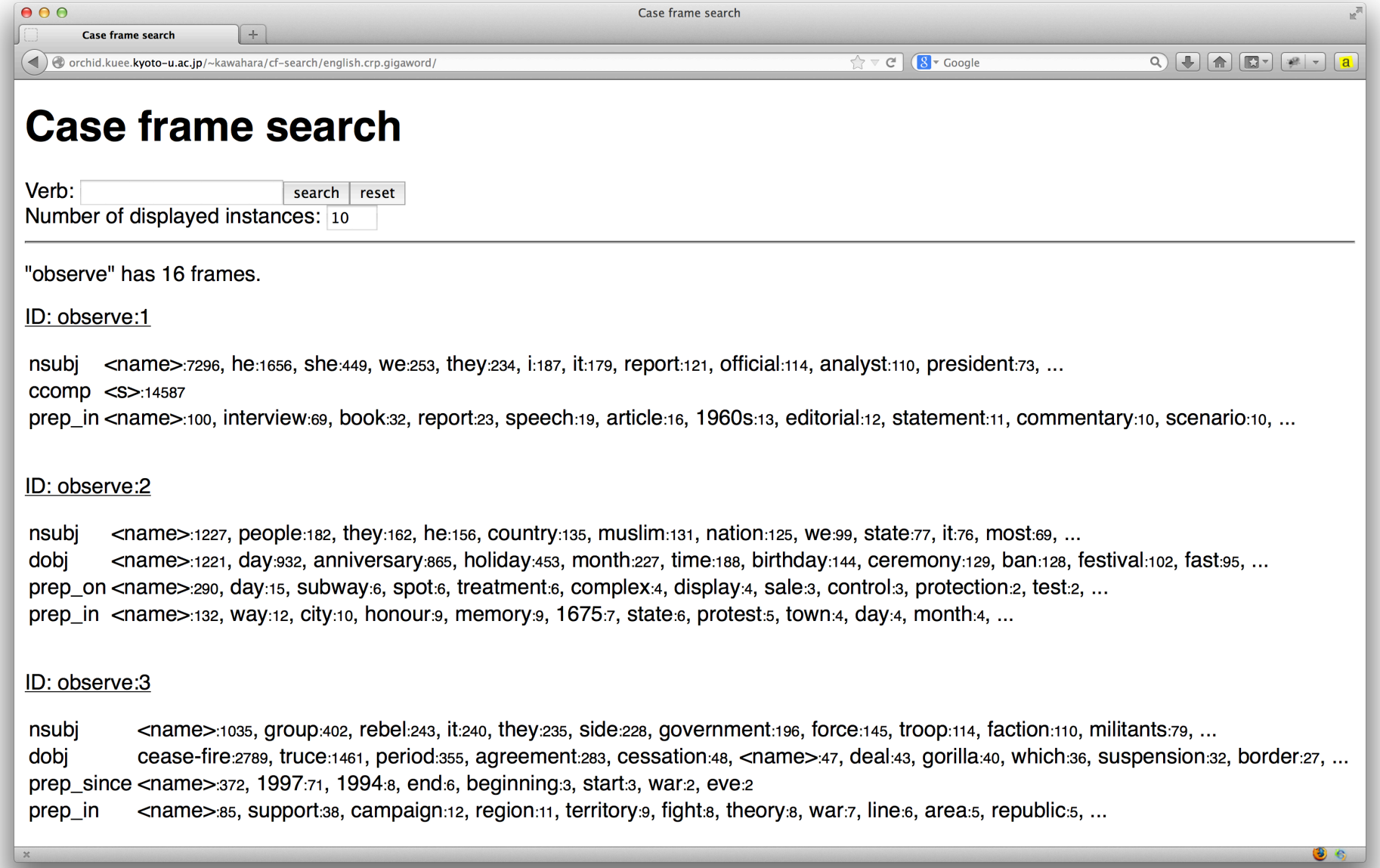
prep_at:{time, point, ..}

prep_at.time:164
prep_at.point:57

nsubj:{you, child, ..} observe dobj:{bird}

nsubj.you:704
nsubj.child:563

dobj.bird:1,692



※ 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