述語項構造と照応関係のアノテーション: NAISTテキストコーパス構築の経験から

飯田龍(NICT), 小町守(首都大),井之上直也(デンソー・東北大),
乾健太郎(東北大), 松本裕治(NAIST)

言語処理学会20周年記念シンポジウム@東大 (2014/10/4)

述語項構造・照応解析

- 「誰が何をどうした」という命題を計算機で扱いやすくする処理
  - 入力) 政府は低所得者を支援する計画を発表した。
    関係省庁の協力を要請する。

	ガ格	ヲ格	ニ格
支援する	政府	低所得者	
発表する	政府	計画	
要請する	政府	協力	関係省庁

  - 日本語は述語の格要素の省略が頻出するため，省略の指し先の特定（ゼロ照応解析）が重要

照応・共参照関係

- 照応関係
  - 文章中である表現が別の表現を指す関係
- 共参照関係
  - 文章中で二つの表現が同一の実体を指す関係

- 例
  - 照応 / 非共参照（Identity-of-sense-anaphora）
    - 太郎はプリウスを買った。次郎も{それを,φを}買った。
  - 照応 / 共参照（Identity-of-reference-anaphora）
    - 太郎はプリウスを買った。次郎は{それに,φに}乗った。

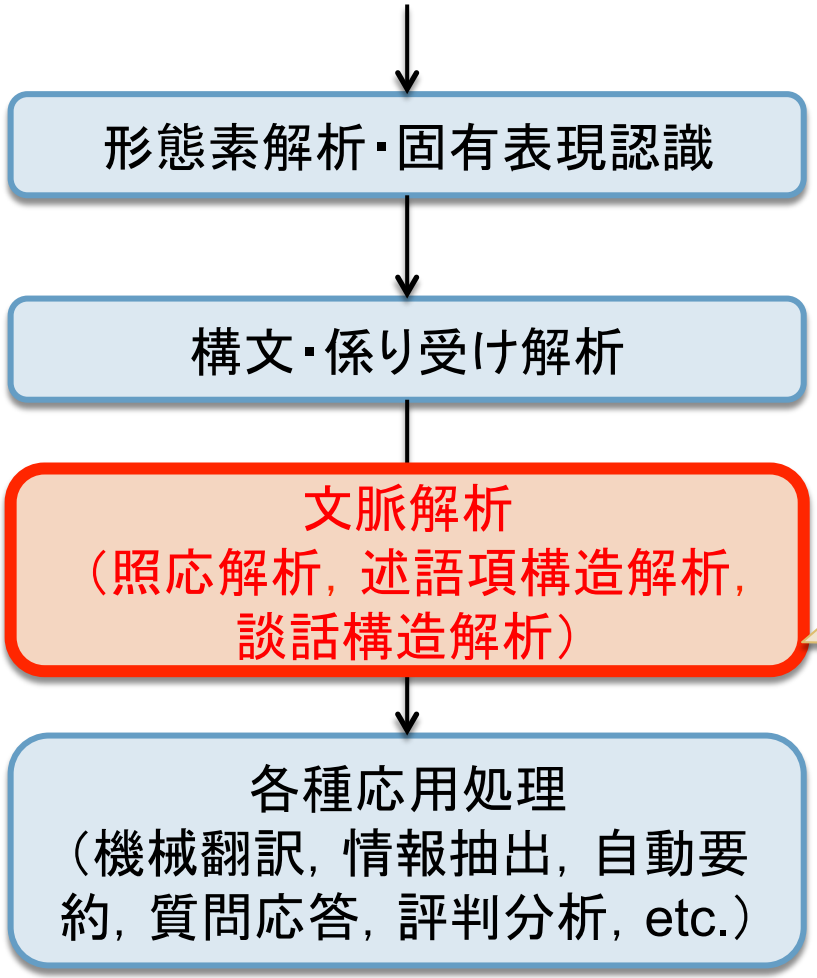
事態性名詞

- 動詞派生の名詞(「走り」,「動き」など)やサ変名詞に対しても述語と同様に項構造を考える
  - 入力) 政府は低所得者を支援する計画を発表した。
    関係省庁の協力を要請する。

	ガ格	ヲ格	ニ格
計画する	政府	支援するコト	
協力する	関係省庁		計画

研究背景

- 現状 / 昔から想定されている処理の流れ



- 構文解析などの表層レベルの処理から一歩理解の問題へ踏み込みたいという動機
- SVMなどの学習アリゴリズムの高度化
- コーパスを整備して，その中の問題を解くことに関する受容

→ 文脈解析の問題をコーパスに基づく研究として進める基盤が整備された

関連研究（内外の動向）

- 述語項構造・意味役割付与
  - PropBank + VerbNet, NomBank
  - GDA, 京都大学テキストコーパス4.0
- 共参照解析
  - Message Understanding Conference 6,7
  - Automatic Content Extraction 2002 – 2007
  - OntoNotes → CoNLL 2011 shared task
  - GDA, 京都大学テキストコーパス4.0

ねらい

- 述語項構造・共参照関係タグ付きコーパスに基づく日本語固有の照応・省略解析器の開発

- その当時の状況
  - 日本語の照応・共参照解析の研究は，研究者が個別にデータを小規模に作成して，それを使ってそれぞれ研究．データは共有されない

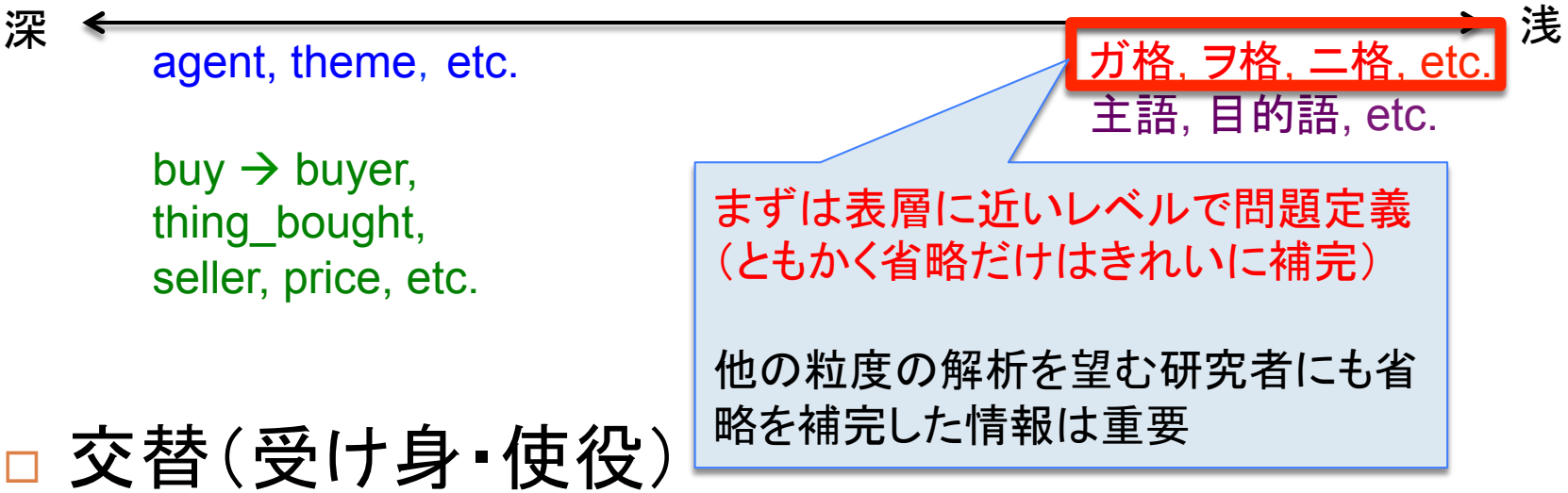
→ 研究者横断的に共有できる述語項構造・共参照関係のアノテーション仕様の作成，それに基づく大規模コーパスの構築

アノテーションスキーマを考える上で何が問題だったのか?

- 述語と項の関係
  - どこまで深く解析する問題とするのか?
  - 交替など表層格が変わる場合は?
  - 何が省略の現象なのか?
  - 必須格をどう捉えるのか?
- (代)名詞句間の照応・共参照関係
  - 照応と共参照どちらを採用するのか?

工夫・議論の対象となった点（1/5）

- 述語と項のラベル



- 交替（受け身・使役）
  - 私は 彼に リンゴを 食べさせる
    - 食べさせる: 私ガ 彼ニ リンゴヲ
    - 食べる: 彼ガ リンゴヲ (使役者:私)

情報抽出を意識した規格化

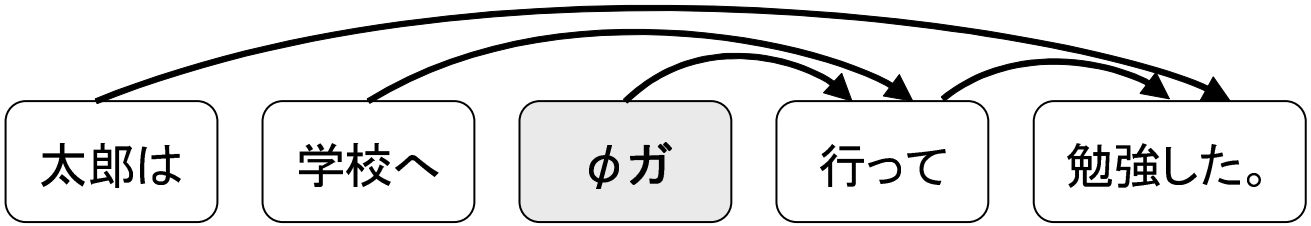
工夫・議論の対象となった点(2/5)

- 何が省略なのか?
  - 人間が省略を検出してアノテーションを行うため，この部分の取り決めが重要
  - 何が任意格で何が必須格なのか?
    - 例) 私は花子に靴を買った。(人ガ モノヲ 買う)
    - cf. PropBankは動詞辞書VerbNetに記載された構文パタンを参照してデータ作成
    - 日本語の辞書：語彙大系構文体系，動詞項構造シソーラス → 網羅性・実際の出現と定義との異なりの点で導入を取り止め

→ 人間の内省で任意・必須の判定
(もちろん辞書も参照)

工夫・議論の対象となった点（3/5）

- 何が省略なのか? (Cont'd)
  - 文内の省略:「太郎は学校へ行って勉強した。」
    - 並列構造? 項の省略?



→ 係り受け構造に基づいて省略を定義

  - 外界照応:指し先が文章内に存在しない
    - 「一人称」,「二人称」,「それ以外」で付ける
    - 「おなかが減ったので、(φ=〈一人称〉ガ)帰ろうと(φ=〈一人称〉ガ)思う。」

工夫・議論の対象となった点(4/5)

- 照応vs共参照
  - 言語的な機能で「指す」関係 or 世界における同一性
    - 日本語は冠詞が無いので,「指す」関係を捉えづらい

    → 共参照関係をアノテーションする
- どこまで共参照と考える?
  - e.g. 総称名詞

本$_a$は、書物の一種で、印刷・製本された出版物を指す。
図書館の本$_b$は借りることができる。

本$_a$ ⊃ 本$_b$

→ 総称名詞間の共参照関係はアノテーションしない

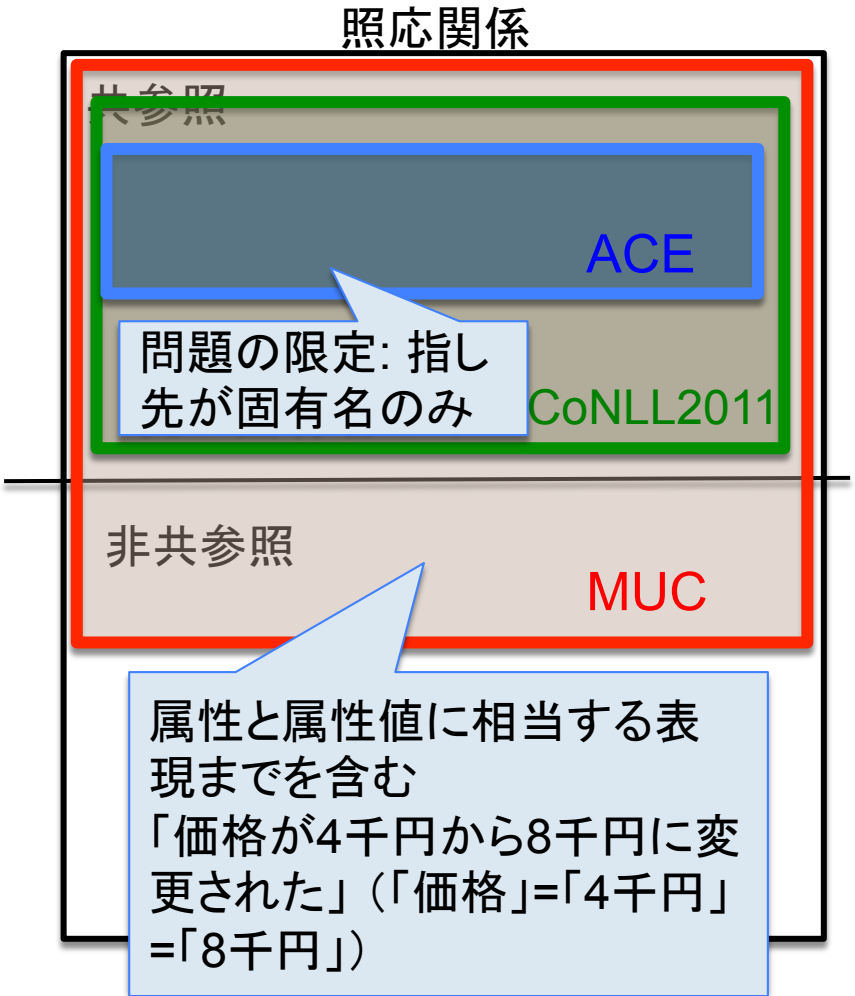
工夫・議論の対象となった点(5/5)

- 共参照関係だけアノテーションすると代名詞・指示連体詞がアノテーション対象外となる
- 指示連体詞の指し方が2種類
  - 指定指示(限定指示)
    - 図書館で資料を手に入れた。このデータは機械的に処理される。(資料 ← このデータ)
  - 代行指示
    - 5年間、水質調査を行った。このデータは機械的に処理される。(水質調査 ← この)

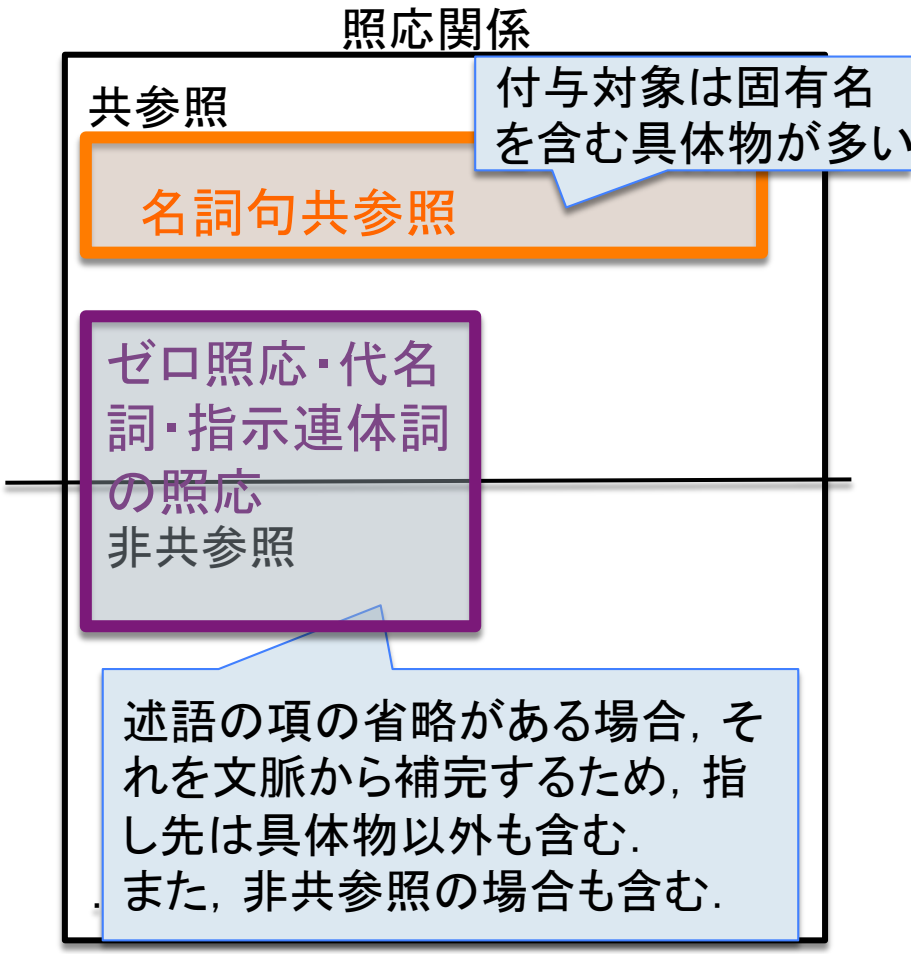
→ これら2つの関係を付け分ける

既存コーパスの問題の異なり

- 英語



- 日本語 (NAISTテキストコーパス)



アノテーションのサイクル

2人～4名(?)毎日
アノテーション作業

グループ会議

飯田

作業・ツールの質問
仕様とのずれの指摘
仕様に書かれていない現象
についてどうするのか?

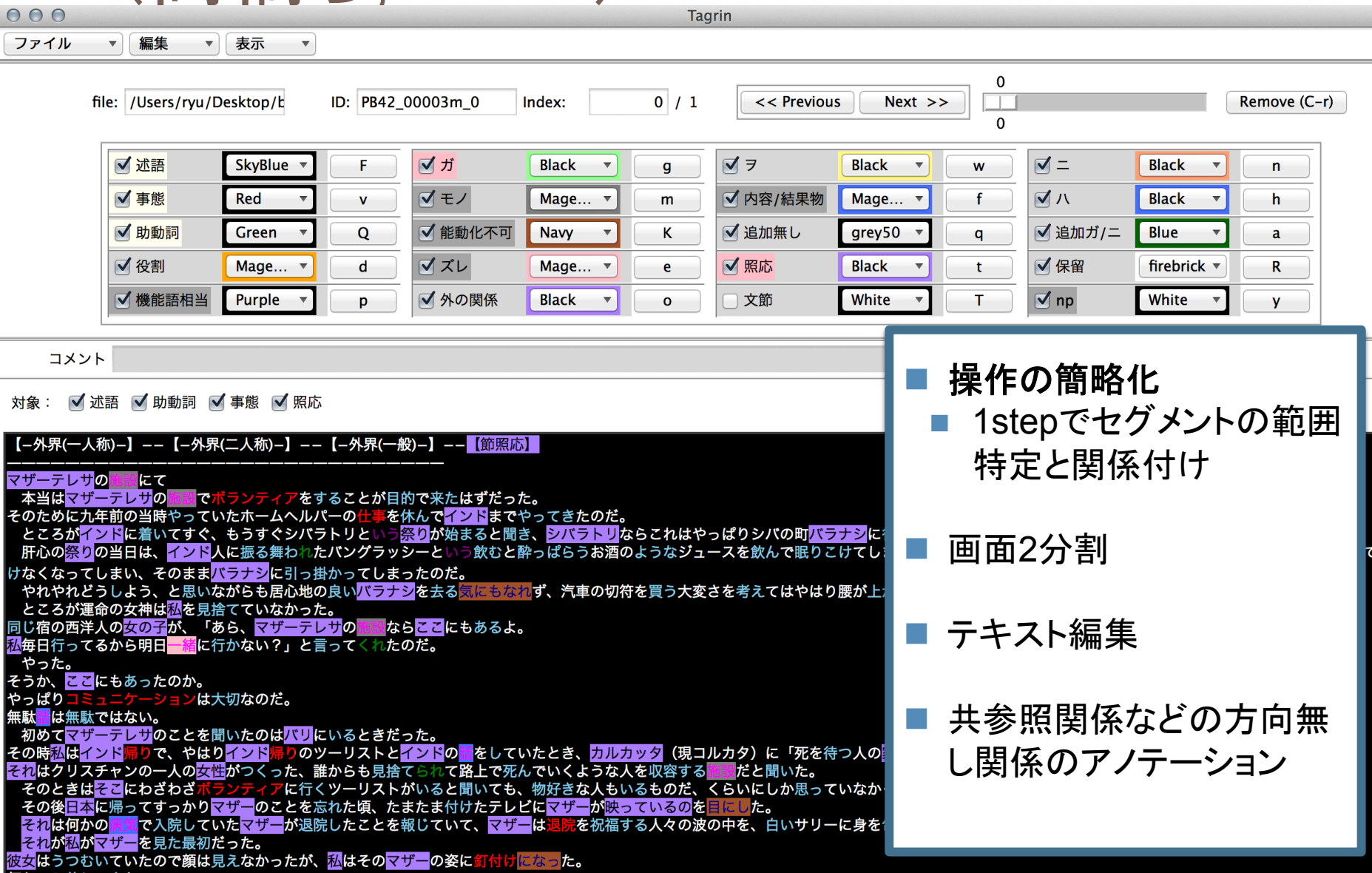
わかる範囲で回答
仕様・ガイドラインの更新

飯田

整合性のとれた基準を仕
様・ガイドラインに反映

未解決の問題を提示

アノテーションツール: Tagrin（高橋ら, 2006）



- **操作の簡略化**
  - 1stepでセグメントの範囲特定と関係付け
- 画面2分割
- テキスト編集
- 共参照関係などの方向無し関係のアノテーション

挑戦（当時）

- 世界最大規模の述語項構造・照応関係タグ付きコーパスを作成

言語	コーパス	語数 / 形態素数
英語	MUC7	29K
	ACE-2007	315K
	OnteNotes	1445K
日本語	NAISTテキストコーパス	1057K
	BCCWJコア(短単位) (作業中...)	1290K

統計量

□ 京都[illegible]ス3.0 (2,929文[illegible]に

約6割が
ゼロ照応関係にある

8割以上が
係り関係にある

約8割が
ゼロ照応の関係

同一文節内に
最も多く出現

		ガ格	[illegible]	[illegible]格
述語	同一文節内	177 (0.002)	60 ([illegible]001)	[illegible]1 (0.027)
106,628	係り関係	44,402 (0.419)	35,882 (0.835)	18,912 (0.879)
	ゼロ照応(文内)	32,270 (0.305)	5,625 (0.131)	1,417 (0.066)
	ゼロ照応(文間)	13,181 (0.124)	[illegible]	[illegible] (0.025)
		15,885 (0.150)	[illegible]	[illegible] (0.002)
		105,915 (1.000)	42,9[illegible]000)	[illegible]07 (1.000)
事態性名詞	同一文節内	2,195 (0.077)	5,574 (0.506)	846 (0.436)
28,569	係り関係	4,332 (0.152)	2,890 (0.263)	298 (0.154)
	ゼロ照応(文内)	9,222 (0.324)	1,645 (0.149)	586 (0.302)
	ゼロ照応(文間)	5,190 (0.183)	854 (0.078)	201 (0.104)
	ゼロ照応(文章外)	7,525 (0.264)	42 (0.004)	10 (0.005)
	全体	28,464 (1.000)	11,005 (1.000)	1,941 (1.000)

一致率

- 作業者2人が30記事をアノテーション

一人の作業結果を正解，もう一人の結果をシステムの出力とみなして，再現率・精度で評価

	再現率	精度
述語	0.921 (806/875)	0.944 (806/854)
ガ格	0.823 (683/830)	0.829 (683/824)
ヲ格	0.899 (329/366)	0.954 (329/345)
ニ格	0.724 (105/145)	0.890 (105/118)
事態性名詞	0.965 (247/256)	0.792 (247/312)
ガ格	0.735 (191/260)	0.743 (191/257)
ヲ格	0.827 (86/104)	0.869 (86/99)
ニ格	0.389 (7/18)	0.583 (7/12)
共参照	0.813 (126/155)	0.813 (126/155)

ご利用いただきありがとうございます

- NAISTテキストコーパス 1.5
  http://cl.naist.jp/nldata/corpus/

  - **公開後約450ダウンロード**
  - **このコーパスを利用した研究が国際会議にも採択されている**
    - Taira et al. (EMNLP2008), Taira et al. (ACL2010), Sasano&Kurohashi (IJCNLP2011), Yoshikawa et al. (IJCNLP2011), Imamura et al. (ACL2009), Hayashibe et al. (IJCNLP2011), etc.

未解決の問題

- 最終的にはアノテータが判断できるかでデータの質が変わる
  - 機能語相当：e.g.「～として」
    - どの表現を機能語的だと考えるかには揺れがある
    - 列挙も難しい・曖昧性もある
  - 名詞句共参照
    - 事態と事態の同一性の判定
    - 抽象名詞の共参照関係?
  - 格パタンの異なり・二重主語構文
    - 欠落している格を人が気付けるのか?

→ 機械的にサポートできるものはやるという体制の整備

今後の方向性

- 現象の多様性はカバーできているのか?
  - SNS文書，論文など別の種類の文書へアノテーションが必要
- 学習アルゴリズム・解析アルゴリズム・特徴抽出のさらなる改善
  - 現象の多様性を捉えるには? 今までの延長線上でうまくいくのか? 問題の観点を効果的に捉えるには?
- 情報共有
  - アノテーション時の経験則，問題の評価指標
- ジレンマ
  - 学術的には問題を安定させなければ共有が難しい / 今後想定される応用に向けて問題の改編が必要
- アノテーションしたデータだけ対象にしていていいのか?
  - **アノテーション学 vs ポスト経験主義**