

日本語コーパスに対する述語項構造のアノテーション

小町研究室（自然言語処理）准教授 小町守 / 協力: 東工大 飯田龍, 東北大 乾健太郎, 奈良先端大 松本裕治

概要

日本語コーパスに対し、照応と述語項構造のアノテーションを行なっている。

- ① 新聞・書籍・白書・知恵袋のコアデータに対し、照応と述語項構造のアノテーションをした
- ② 述語に関する辞書を参照してアノテーションを行い、質の向上に関する有効性を確認した

研究背景

背景

✓言語理解のためには述語項構造解析が必要

✓述語項構造解析は情報抽出や文書要約などに使える

問題点

① 新聞記事に対するタグつきコーパスしか存在しない

② 複数人でのアノテーションの一致率が低い

目的

① 多様なジャンルの述語項構造タグ付きコーパスの作成

② 人手によるアノテーションの質の向上

照応・述語項構造解析

関連研究

✓京都テキストコーパス 4.0 (Kawahara et al., 2002)

✓KNB コーパス (橋本ら, 2009): ブログコーパス

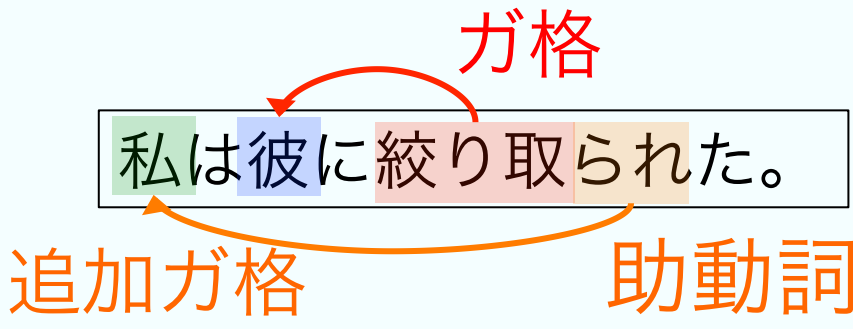
✓GDA コーパス (Hasida, 2005)

✓NAIST テキストコーパス (飯田ら, 2010)

アノテーション方針と進捗

述語項構造のアノテーション

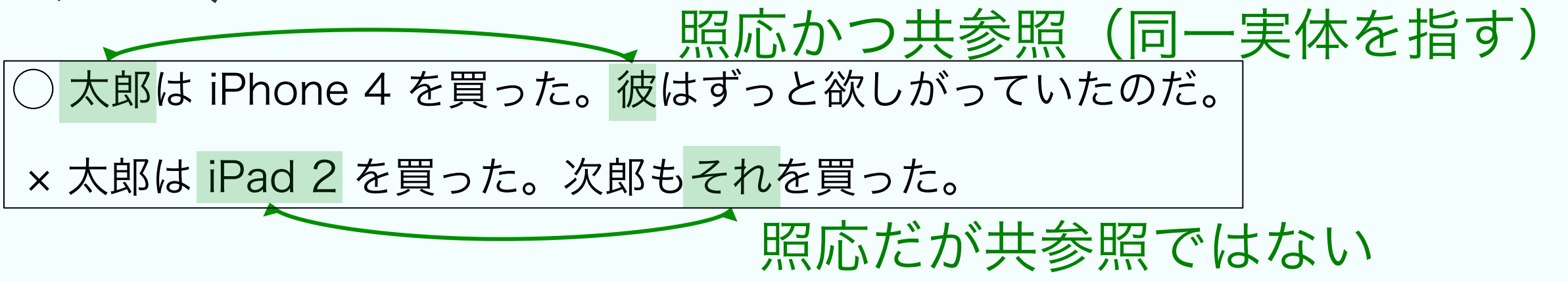
述語項構造については、述語の基本形に必須格となる表現を表層格（ガ・ヲ・ニ格）レベルでタグ付けする。その際、格交替は原形に戻す。事態性名詞についても、述語と同様に表層格レベルで項を付与する。
(必須格の判定には事態間関係知識 (青山ら, 2007; 大西ら, 2008) 語彙概念構造辞書 (竹内, 2004) を使用)



助動詞として扱う語のリスト		
	パターン	例
補助動詞	てあげる、てもらう、...	代筆してあげる、食べてもらう
可能	できる、れる、られる	説得できる、行かれる、見られる
受け身	れる、られる	騙される、認められる
使役	せる、させる	喋らせる、論じさせる
願望	たい	食べたい

照応関係のアノテーション

照応かつ共参照関係にある場合のみ照応関係を認定する。
(NAISTコーパスと異なり、bridging reference および間接照応 (Inoue et al., 2010) の情報は付与していない)



アノテーションの進捗

- ✓ 4ジャンルでコアデータAまで完了
- ✓ 雑誌・ブログデータは未着手

コアデータのタグ付け進捗			
	記事数	文数	単語数
新聞	478	5,730	127,077
書籍	55	4,691	113,399
白書	30	2,414	100,396
知恵袋	938	6,402	103,188

議論

コーパスの質の向上
コーパスアノテーション支援ツール Slate (徳永ら, 2010) を用いて一致率の向上を目指す (cf. OntoNotes (Hovy et al., 2006))

事態性名詞に関するリソースの構築
NOMLEX (Macleod et al., 1998) や NomBank (Meyers et al., 2004) のように名詞に関するリソースを作成

多層的なアノテーション
述語の語義付与・意味役割付与とそれらの同時自動推定手法の開発

コーパスの扱いやすさの向上
CoNLL 形式など扱いやすいフォーマットでリリースする

コーパスを用いた自動解析手法の研究
新聞記事で訓練された述語項構造解析モデルの分野適応

作業者間の新聞コアデータにおける一致率

作業者A-B間			作業者A-C間		
	適合率	再現率	適合率	再現率	
述語	87.1 (155/178)	89.6 (155/173)	80.1 (133/166)	76.9 (133/173)	
ガ格	80.4 (123/153)	100.0 (123/123)	96.2 (101/105)	100.0 (101/101)	
ヲ格	96.3 (77/80)	98.7 (77/80)	85.3 (58/68)	96.7 (58/60)	
ニ格	82.1 (23/38)	88.5 (23/26)	90.5 (19/21)	59.4 (19/32)	
事態	93.9 (93/99)	84.5 (93/110)	72.9 (70/96)	63.6 (70/110)	
ガ格	83.5 (76/91)	100.0 (76/76)	80.0 (32/40)	100.0 (32/32)	
ヲ格	59.0 (13/22)	100.0 (13/13)	52.4 (11/21)	73.3 (11/15)	
ニ格	77.8 (7/9)	100.0 (7/7)	50.0 (4/8)	100.0 (4/4)	

- 対象: 新聞記事9記事 (90文・1,653語)
評価: 一方の作業者のタグを正解とし、他方の作業者のタグを出力とした適合率と再現率
作業者: A=NAISTコーパスのアノテーションに従事した熟練作業者・B,C=自然言語処理が専門の大学院生
- ✓ 辞書を参照するべきかどうか、作業者による揺れ
 - ✓ 事態性名詞に関する辞書がない