文献紹介

自然言語処理研究室
後藤　大明

出典

* A Fully Coreference-annotated Corpus of Scholarly Papers from the ACL Anthology

* Ulrich Schäfer ,Christian Spurk,Jörg Steffen,COLING 2012 pp.1059-1069

背景

* 同一指示関係(CR)について
* 指示：代名詞の指示する内容が言語外に存在する
* 照応：他の言語表現との同一の指示対象を表す
* これまでのCRの研究は新聞記事などの領域で行われてきた

背景

* 科学文章のCRは新聞記事に比べ困難、相対的に合成的で複雑な傾向がある[Watson et al., 2003]

* 科学文章に対応した新たなコーパスを作成する

先攻研究

* 生物医学テキスト内の照応と同一指示の完全な注釈を提示[Gasperin, 2009]
* 生物学テキスト97全文雑誌論文からすべて共同参照名詞句が注釈されているコーパスを構築[Cohen,2001]

コーパスの構築

* ACLアンソロジーから266の論文を用いて作成
* テキストは商用のOCR機能を用いてPDFから抽出
* 抽出の品質は、PDFジェネレータに依存

アノテーション

* アノテータは一般的な言語学の知識を持つ学生
* ACE のアノテーションタスクと同様に行う
* 48960の文章で1326147のトークンを付与

アノテーション結果

Mention Type	Amount
`def-np` (definite NPs)	32,547
`pper` (personal pronouns)	5,921
`ne` (proper names incl. citations)	14,451
`ppos` (possessive pronouns/determiners)	3,407
`indef-np` (indefinite NPs)	6,820
`conj-np` (coordinations)	1,446
`pds` (demonstrative pronouns)	435
`prefl` (reflexive pronouns)	266
Σ	65,293

アノテーション結果について

* Definite Noun Phrases:個人を特定できるエンティティに対応する NP “the”,”that”

* Indefinite Noun Phrases:指定されたコンテキストで特定し、識別可能なエンティティに対応しないNP

* Personal Pronouns:人称代名詞　“I”,”you”

誤差解析、修正

* コーパスの13%が異なるアノテータによってによって2回アノテーション
* アノテーター達での意見の一致
* 12%のアノーテション修正

おわりに

* ACLアンソロジーに採録された266の論文を用いて同一指示タグつきコーパスを作成し公開した
* 機械学習などのトレーニングデータのみでなく他のNLPタスクへの貢献が期待できる