構文解析器開発実験の観点から見たコーパスアノテーションの揺れの分析

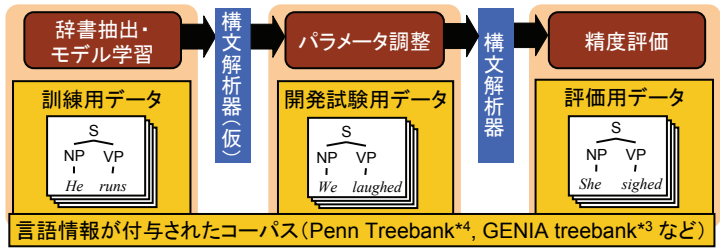
原 忠義[1] 辻井潤一[1,2,3]
[1] 東京大学情報理工系研究科コンピュータ科学専攻辻井研究室, [2] School of Computer Science, University of Manchester,
[3] NaCTeM (National Center for Text Mining)

大規模コーパスを用いた構文解析器開発実験

各実験過程において, 言語情報が付与されたコーパスを利用

- 大量の実データに基づく頑健な開発実験 → 高い解析精度を達成
- コーパスのアノテーションの質に依存
  - Enju*1*2 の解析エラー中 **23.2 %**は評価用データの「不具合」による（GENIA treebank*3 50文に対するエラー166個を人手で分析）

→ アノテーションの「不具合」と それらの実験への影響を調査



コーパスアノテーションの揺れ

注目するのは個々のケアレスミス等ではなく、全体的な「揺れ」

(PP *due to* (NP *this*)) : *53箇所* ⇔ (ADVP *due* (PP *to* (NP *this*))) : *53箇所*

- モデル学習・パラメータ調整 → 曖昧性が高くなり, 不安定に
- 精度評価 → 評価が不当に上昇・下降

多種多様なアノテーションの揺れが存在 (GENIA treebankの例)

前置詞句に対するアノテーションの有無 ／ coordination vs. conjunction ／ ハイフン分割 の有無／ coordination の構造・範囲 ／ 副詞の係り先 ／ 括弧のアノテーションの仕方 ／名詞句に対する定冠詞の位置付け ／ 他多数

→ 数種の「揺れ」について, それらの実験への影響を確かめる

アノテーション揺れの修正実験

手作業で GENIA treebank (18,541文) 中の「揺れ」数種を修正

アノテーション揺れの種類	修正箇所
Coordination vs. conjunction	176 箇所
名詞句中の "*of*" のアノテーション方針	208 箇所
言い回し表現中の "*of*" のアノテーション方針	117 箇所

修正した揺れ(1): Coordination vs. Conjunction

例: *A induced C, whereas B did not.*

- **Coordination の "whereas": 142 箇所**
  (S-COOD (S (NP *A*) (VP-1 *induced* (NP *C*))) *, whereas* (S (NP *B*) (VP *did not* (VP *?*-1))) *.*/.)
- **Conjunctive の "whereas": 148 箇所**
  (S (NP *A*) (VP-1 *induced* (NP *C*)) *,* (SBAR-ADV *whereas* (S (NP *B*) (VP *did not* (VP *?*-1)))) *.*/.)

統一

その他の対象箇所: "*i.e.*" など

修正した揺れ(2): 名詞句中の "*of*" のアノテーション方針

例: *nuclear factor(s) of activated T cell(s)*

- **フラットな構造: 63 箇所**
  (NP *nuclear factor of activated T cell*)
- **詳細な構造が付与: 40 箇所**
  (NP (NP *nuclear factor*) (PP *of* (NP *activated T cell*)))

統一

その他の対象箇所: "*signal transducer and activator of transcription* (*STAT*)", "*short arm of chromosome*", "*down-regulation of inflammation*", *他多数*

修正した揺れ(3): 言い回し表現中の "*of*" のアノテーション方針

例: *A occurred in the presence of B*

- **フラットな構造: 44 箇所**
  (S (NP *A*) (VP *occurred* (PP ***in the presence of*** (NP *B*))))
- **詳細な構造が付与: 72 箇所**
  (S (NP *A*) (VP *occurred* (PP ***in*** (NP (NP ***the presence***) (PP ***of*** (NP *B*))))))

統一

その他の対象箇所: "*in the absence of*", "*by virtue of*", "*in spite of*", *他多数*

修正コーパスでの構文解析実験

オリジナルコーパスと修正コーパスで実験結果を比較

- 構文解析器: Enju*1
  - Penn Treebank*4 02-21節 (39,832 文) で学習後,
  - GENIA treebank*3 (14,849 文) で追加学習, 生医学分野に適応*2
- 修正された GENIA treebank で Enju を再学習・再評価
  - 14,849 文で学習, 1,850 文で開発テスト, 1,842 文で評価

GENIA treebank	F-score (LP / LR)	Diff.
オリジナル	88.39 (88.70 / 88.08)	
揺れを修正	88.92 (89.24 / 88.61)	+ 0.53

「言い回し表現中の "*of*"」に関する修正効果

対象箇所	修正前 モデル学習	修正前 評価	修正後 モデル学習	修正後 評価
in the presence of	25 vs. 71	19 vs. 1	0 vs. 96	0 vs. 20
in the absence of	29 vs. 56	4 vs. 0	0 vs. 85	0 vs. 4

(フラットな構造の数 vs. 詳細構造の数)

- 構文解析精度: +0.36 (F-score)
- 修正前: 全体的な揺れ, モデル学習と評価で正しい構造が逆転
  → 修正により、各データセットの一貫性・学習と評価間の整合性

→ コーパスの「揺れ」による実験への影響が無視できない割合に

実験側の要求 vs. コーパス側の方針

コーパスの「揺れ」が必ずしも「不具合」とは限らない

- 構文解析実験側から見て不都合なだけのケースも
  - 「揺れ」をコーパスがバリエーションとして認めている
  - 「揺れ」の修正方針が実験側とコーパス側で真逆に

→ 多様な「揺れ」を適切に処理する必要性

例: 専門用語内の構造は省かれ易い (GENIA treebank)

アノテーション: (NP *heat shock protein 90 complexes*)

省略された内部構造: *heat shock* | *protein* | *90* | *complexes*

→ アノテーションの不備ではなく, コーパス側の方針

より高品質な構文解析器の開発へ向けて

- アノテーションの揺れを徹底的に洗い出しその全貌を解明する → 必要であればコーパスへのフィードバック
- アノテーションの揺れを前提とした構文解析器開発実験手法の模索
- 異なるコーパス間での「揺れ」による影響の検証 → 分野適応など, 複数コーパス利用時の精度向上手法へ

参考文献:
[1] T. Ninomiya et al. 2006. Extremely lexicalized models for accurate and fast HPSG parsing. In Proceedings of 2006 EMNLP, pages 155-163.
[2] T. Hara et al. 2007. Evaluating Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser. In the Proceedings of IWPT 2007.
[3] Kim et al. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl.1). pp. i180-i182.
[4] M. Marcus et al. 1994. The Penn Treebank: Annotating predicate argument structure. In Proceedings of ARPA Human Language Technology Workshop.