E9-5

Annotation of modal expressions in Indonesian

HIROKI NOMOTO JOZINA VANDER KLOK
DAVID MOELJADI

東京外国語大学
Tokyo University of Foreign Studies

はじめに

- インドネシア語を対象に、モダリティに関連する情報を付与したデータセットを構築
- https://github.com/matbahasa/IndoModal にて公開
- 目的：言語学的研究
- 本発表
  - データセット構築の方法
  - 構築したデータセットから分かった言語事実

テーマセッション「深層学習時代の言語学と自然言語処理」との関係

- 「深層学習／自然言語処理と言語学を融合させた研究の発表の場を提供することを目的とする」
  →そういう研究ではありません
- 「深層学習や自然言語処理によるサポートが有益となる場面も少なくない」
  →どのようなサポートが可能か知りたい・考えてもらいたい
- （ここにはほとんどいない）伝統的な言語研究をやっている言語学者、日英語以外を対象とする言語学者がやりたいことを伝え、興味を持ってもらう

背景

・モダリティとは

・インドネシア語のモダリティ

モダリティとは

- 必然性・可能性を表す意味カテゴリー
- モダリティ表現の例

  must, should, can, はずだ, べきだ, かもしれない
- モダリティの意味の構成要素
  - **強さ（force, strength）**

    可能性（possibility）、必然性（necessity）、弱必然性（weak necessity）
  - **種類（flavour）**

    認識（epistemic）、根源（root）

モダリティの意味の構成要素①：強さ

- **可能性**

  Ann **may** buy a lottery ticket because she turned 18 years old.

  （アンは18歳になったので宝くじを買ってもよい。）

- **必然性**

  Ann **must** buy a lottery ticket because her boss ordered her to.

  （アンは上司に命令されたので宝くじを買わなければならない。）

- **弱必然性**

  Ann **ought to** buy a lottery ticket today if the odds are good.

  （アンは当たる可能性が高いなら今日宝くじを買うべきだ。）

モダリティの意味の構成要素②：種類

- **認識**：話者の知識から見た必然性・可能性

  Ann **may** buy a lottery ticket today because she's feeling lucky.

  （アンは気分がいいので、今日宝くじを買うかもしれない。）

  ［認識・可能性］

- **根源**：その他の点（規則、現実世界の事実）から見た必然性・可能性

  Ann **may** buy a lottery ticket today because she just turned 18 years old!

  （アンはちょうど18歳になったから、今日から宝くじが買える！）

  ［根源・可能性］

英語のモダリティ表現は強さは区別するが、種類は区別しない

インドネシア語のモダリティ表現

- 本研究ではharusとmestiを扱う（いずれも英語ではmustと訳される）
- 接辞-nyaおよびse-...-nyaによる派生形が存在

語幹形	-nya形	se-...-nya形
harus	harusnya	seharusnya
mesti	mestinya	semestinya

Besok	saya	*harus/seharusnya*	ke	kedutaan.
tomorrow	I	must/should	to	embassy

'I must/should go to the embassy tomorrow.'

コーパスにより明らかにしたい言語事実

1. **語幹形の強さ**：派生形は弱必然性で見解が一致するが、語幹形については
   - 必然性のみ（Sneddon et al. 2010）
   - 必然性・弱必然性（Quinn 2001, Stevens & Schmidgall-Telling 2004, 舟田他 2018）
2. **種類**：英語のモダリティ表現同様、同形が認識にも根源にもなるはずだが、多くの辞書では
   - mestiとsemestinyaは、両方の意味
   - harusとseharusnyaは、根源の意味のみ
3. **法（mood）の関与**：-nya形・se-...-nya形は反実仮想？（Arka 2013）
4. **-nya形の記述**：辞書によってはなし。ある場合も意味記述にばらつきあり

データセット構築の方法

・データ

・アノテーター

・タグセット

データ

- ライプツィヒコーパスコレクション（Goldhahn et al. 2012）のインドネシア語サブコーパスmixed-tufs4, web-tufs13, wikipedia-tufs16
- マレー語・インドネシア語コンコーダンサーMALINDO Conc（Nomoto et al. 2018）を用いて6つのモダリティ形式を含む文を抽出
- 各形式、ランダムに100文を選択

Table 1 Target modal forms and their frequencies

Stem form		*-Nya* form		*Se-. . . -nya* form	
harus	27,245	*harusnya*	224	*seharusnya*	2,057
mesti	524	*mestinya*	314	*semestinya*	247

アノテーション用ファイル（MS Excel）

200143	Rian jahat, kita harusnya gak boleh melakukan ini katanya sambil menangis.	http://17tahun-abay.blogspot.com/
400053	Kadang ide sudah ada, tetapi penulis masih bingung, apa mesti dilakukan dengan ide itu karena begitu gelap untuk menjabarkannya.	http://albasanto.wordpress.com/2009/09/24/membuat-
500167	Biasanya jantung penderita berdetak tidak normal atau tidak berdetak sebagaimana mestinya.	http://ayinosa31.wordpress.com/2010/03/29/olahraga-
106368	Oleh karena itu harus memiliki ilmu manajeman agar segala permaslahan dapat diselesaikan dengan baik, tidak merugikan salah satu pihak dan memuaskan semuanya.	http://apbusinessmanagement.blogspot.com/2009/04/tuga
108427	Juga harus kita akui bahwa pemahaman keliru yang terdapat dalam sebuah buku kadang bersumber dari kesalahan pribadi dari sang penulis tanpa ada maksud jelek dari sang	http://ainuamri.wordpress.com/risalah-jihad-islam-anti-

ID　　文（ランダムに並べ替えてある）　　URL（文脈確認のため）

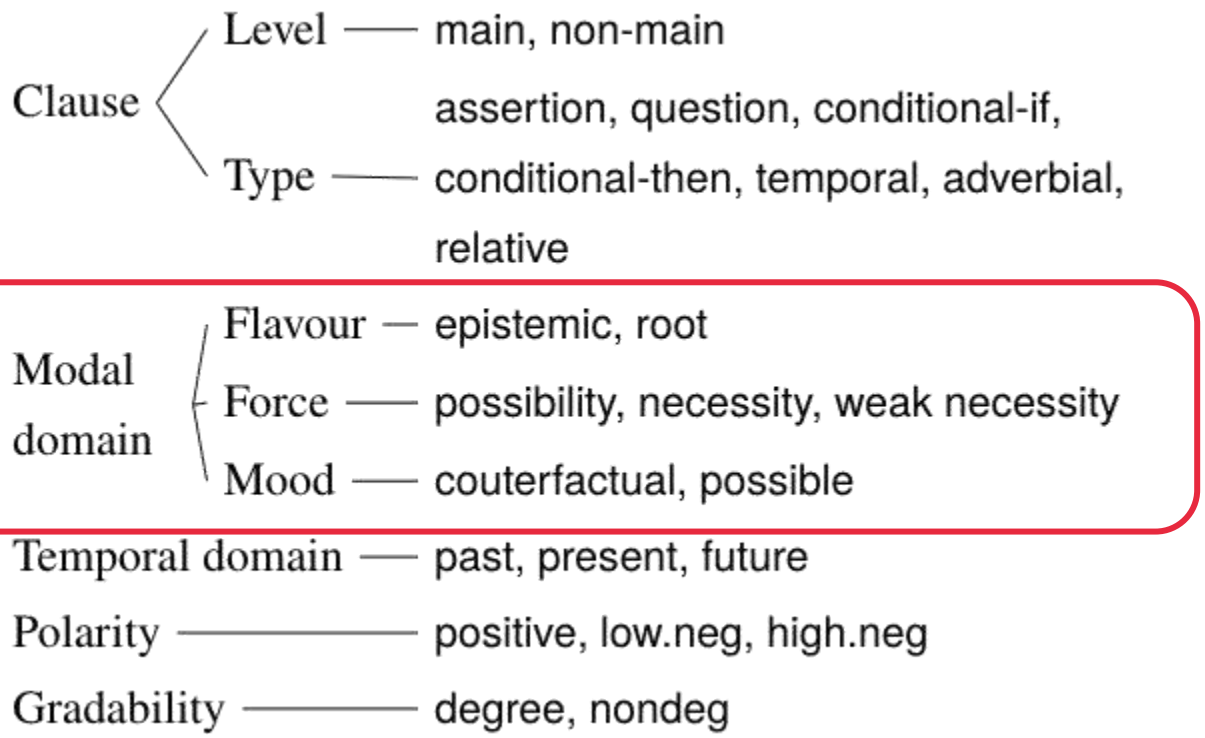
アノテーター

- 3名の母語話者

1. 大学教員、言語学博士号、ジャワ出身
2. 大学教員、言語学修士号、バリ出身
3. 学部生、日本研究専攻、リアウ出身

タグセット

- Rubinstein et al. (2013) とTjuka et al. (2019) を模範に設計



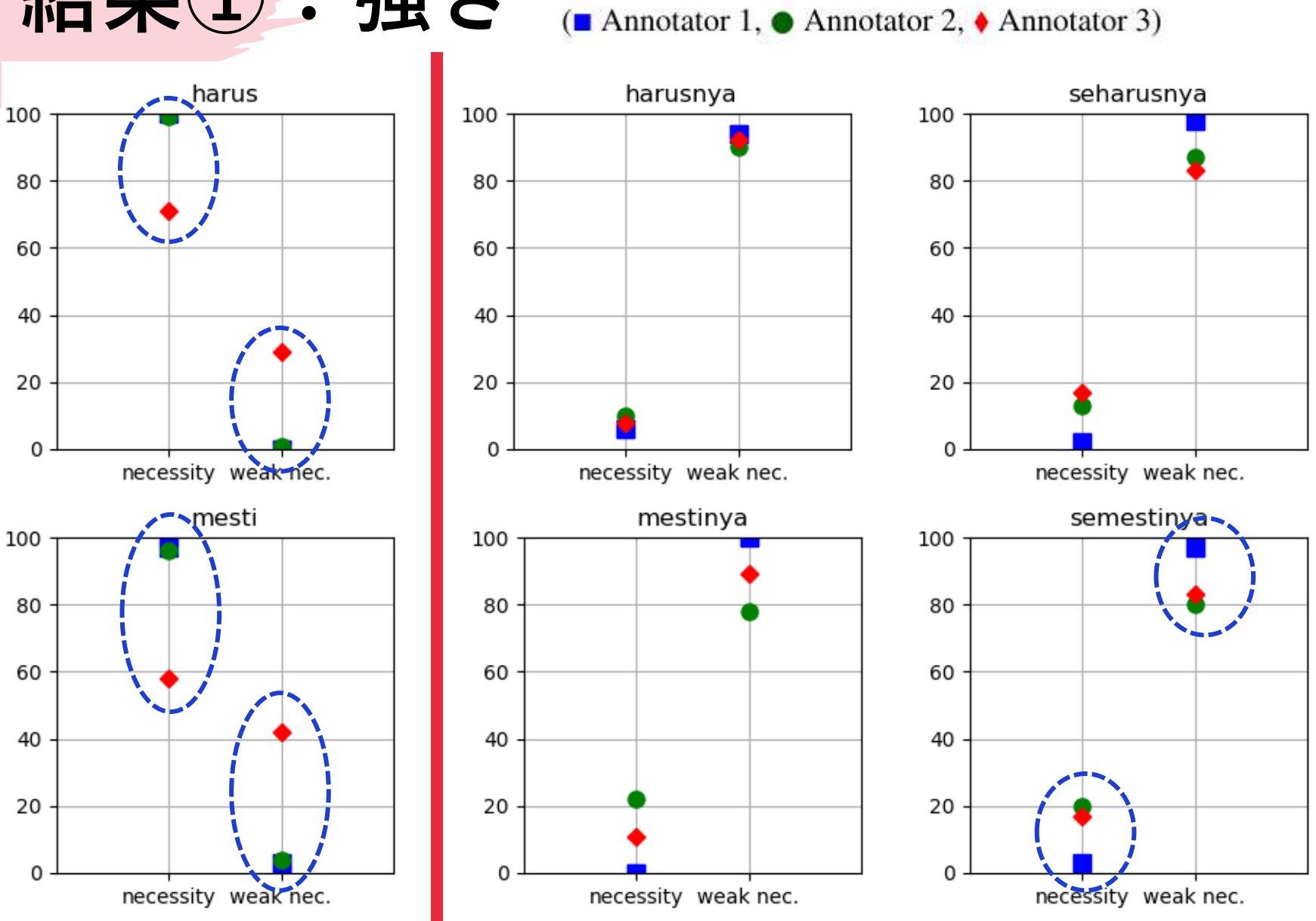
結果と考察

・強さ

・種類

・法の関与

結果①：強さ



「派生形は弱必然性」という先行研究の記述と一致

アノテーター間にばらつきが見られる

アノテーター1, 2
「語幹形は必然性のみ」を支持
アノテーター3
「語幹形は必然性・弱必然性」を支持

アノテーター間一致率

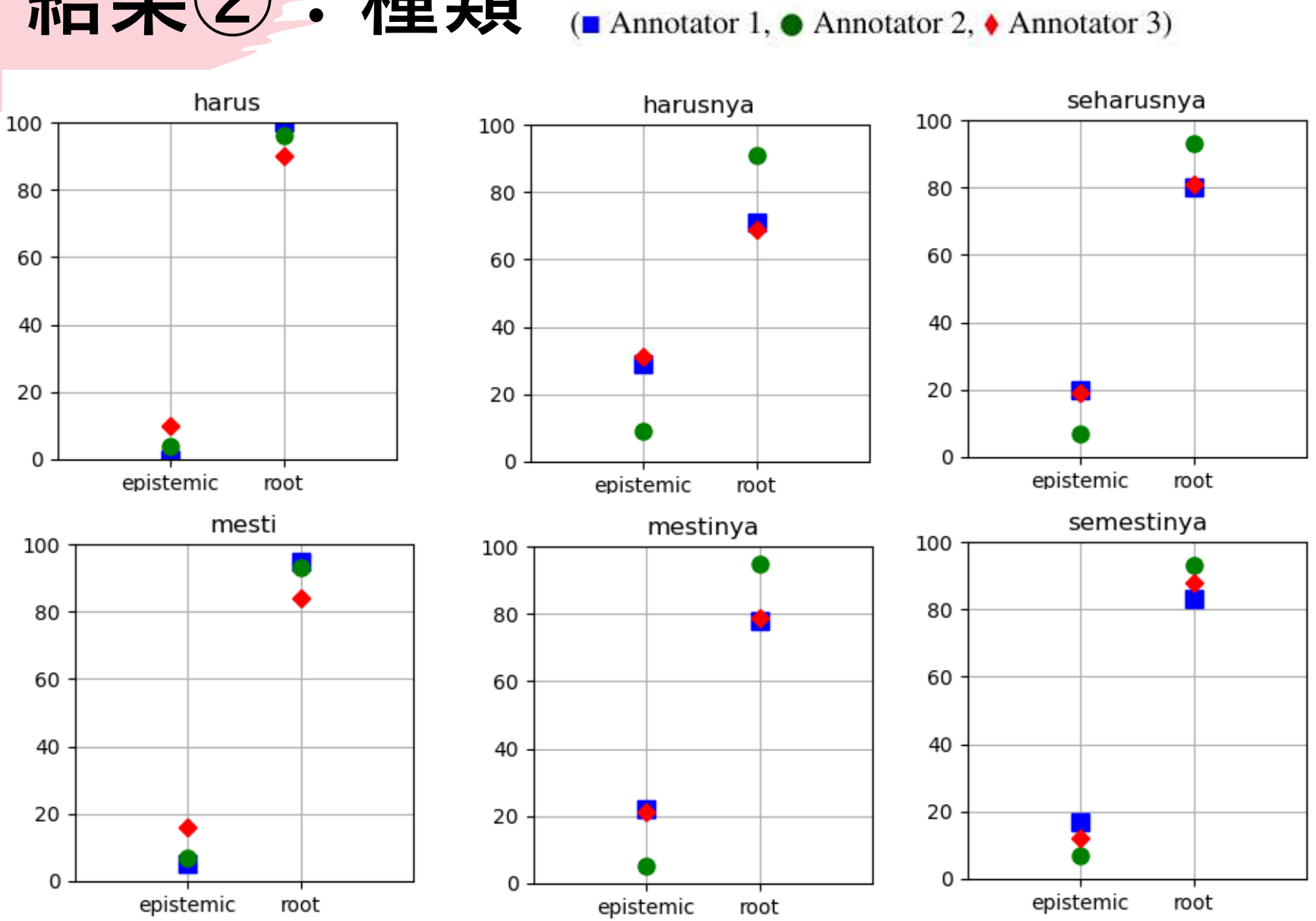
Table 2 Inter-annotator agreement: Modal force

Metric	*Harus*	*H-nya*	*se-H-nya*	*Mesti*	*M-nya*	*se-M-nya*	Total
%	80.0	86.7	90.0	71.3	91.7	76.7	80.3
κ	−0.01	0.09	0.21	0.06	0.15	0.01	0.58

評価指標の問題

- 単純なパーセントは、偶然の一致の可能性が考慮されない
- フライスの κ やクリッペンドルフの α といった標準的指標は、アノテーション対象項目のタグの分布に偏りがある場合、信頼できない（Quarfoot & Levine 2016）

結果②：種類



接辞付加はモダリティの種類に影響を与えない

全体として根源（root）が圧倒的に多い
→標準的評価指標が使えない

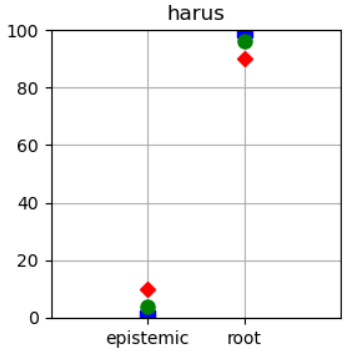
英語の類似研究との比較

Hacquard & Wellwood (2012)

- コーパスを用いて、各種統語環境における認識・根源の分布を調査
- 意味論研究者2名がアノテーション
- must: コーパス全体の89.9%を占める主節で、認識17% vs. 根源 83%
- must, have to: $\kappa = 0.84$

κだけでなく%も報告されていれば、言語間の比較ができた

本研究

- 非意味論研究者3名がアノテーション
- 

両言語とも根源の方が多いが、インドネシア語の方が分布が極端

- 全形式: 79.3%, $\kappa = 0.11$
- harus: 91.3%, $\kappa = 0.04$

認識用法は存在するのか？

- 3名全員がepistemic（認識）とした文が存在
  - harus　1
  - harusnya　6
  - seharusnya　2
  - mesti　2
  - mestinya　2
  - semestinya　1
- harusnya以外は、この数でyesと言ってよい？

結果③：反実仮想

- Arka (2013)
  -nya形・se-...-nya形は反実仮想

a. Kamu *harus* datang.
   you must come
   'You should come.'
b. *Harus-nya* kamu datang.
   must-NYA you come
   'You should have come.' (Arka 2013: (37))

- この記述は明らかに誤り
- 3名全員がpossible（非反実仮想）とした文
  - harusnya 61/100
  - seharusnya 67/100
  - mestinya 53/100
  - semestinya 59/100

結論

・まとめ

・限界

まとめ

- インドネシア語のテキストにモダリティ情報を付与する初めての試み
- 本研究で用いたタグセットとガイドラインは他の研究でも使えるはず
- アノテーション結果は、定性的な先行研究の主張を支持することもあれば、主張に対する反例となることもあった
- インドネシア語ではモダリティの種類において分布に著しい偏りがあることを示した
  - 頻度が低い認識用法の扱いは要注意
  - 標準的なアノテーター間一致率の評価指標より単純な一致率の方が有益

限界

- harus, mestiとその派生形のみ扱ったが、インドネシア語にはより多くのモダリティ表現が存在する
- 600文というデータ
  - サイズ：言語学的研究には十分なサイズだが、NLPのタスクには小さい
  - 形式：文脈なしの文

→文章（ドキュメント）に対するアノテーションが理想的

問題：インドネシア語にはBCCWJ的な汎用コーパスがない

最近多くのデータセットが作られているが、どれも個別のNLPタスク専用の大量の「文＋ラベル」で、言語学的研究には使えそうにない

テーマセッション「深層学習時代の言語学と自然言語処理」との関係（3/11追加）

言語学的問い：

インドネシア語を含む世界の諸言語のモダリティの種類の①個別言語における分布、②言語間の違いの実態はどのようになっているか？

NLPを生かした解法：

1. 機械翻訳で日本語化し、対応する日本語表現の種類（認識・根源）を調べる（英：must、日：に違いない、なければならない）
2. 認識・根源で曖昧にならないパラフレーズ表現との意味類似度を調べる

謝辞

本研究はJSPS科研費 JP23H00639の助成を受けたものです。

参考文献

Arka, I Wayan. 2013. On the typology and syntax of TAM in Indonesian. *NUSA* 55: 23–40.

舟田京子・高殿良博・左藤正範（編著）2018.『プログレッシブ　インドネシア語辞典』小学館.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 759–765.

Hacquard, Valentine & Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics* 5(4): 1–29.

Nomoto, Hiroki Nomoto, Shiro Akasegawa & Asako Shiohara. 2018. Building an open online concordancer for Malay/Indonesian. The 22nd International Symposium on Malay/Indonesian Linguistics (ISMIL)での発表論文.

Quarfoot, David & Richard A. Levine. 2016. How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician* 70(4): 373–384.

Quinn, George. 2001. *The Learner's Dictionary of Today's Indonesian*. Sydney: Allen & Unwin.

Rubinstein, Aynat ,Hillary Harner, Elizabeth Krawczyk,Daniel Simonson, Graham Katz & Paul Portner. 2013. Toward fine-grained annotation of modality in text. In Paul Portner, Aynat Rubinstein & Graham Katz (eds.) *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, 38–46,

Sneddon, James Neil, Alexander K. Adelaar, Dwi N. Djenar & Michael Ewing. 2010. *Indonesian: A Comprehensive Grammar*. London: Routledge, 第二版.

Stevens, Alan M. & A. Ed. Schmidgall-Telling. 2004. *A Comprehensive Indonesian-English Dictionary*. Athens, OH: Ohio University Press.

Tjuka, Annika, Lena Weißmann & Kilu von Prince. 2019. Tagging modality in Oceanic languages of Melanesia. In Annemarie Friedrich, Deniz Zeyrek & Jet Hoek (eds.) *Proceedings of the 13th Linguistic Annotation Workshop*, 65–70.

【参考】Hacquard & Wellwood (2012)

Environment	Total *must*		By flavor				
Total corpus	88,859	100.00%	Epistemic *must*		Root *must*		*p*
Antecedents of conditionals	213	0.24%	1	0.00%	212	0.24%	**
Matrix questions	277	0.31%	34	0.04%	243	0.27%	*
Complements of attitudes	8,034	9.04%	*80*	*0.09%*	*451*	*0.50%*	-
Total embedded	**8,524**	**9.59%**					
Matrix clauses	79,887	89.90%	*68*	*17.00%*	*332*	*83.00%*	

Table 2 **Distribution of epistemic and root *must* in various environments:** ** $p < 0.001$, * $p < 0.05$, - $p > 0.05$, Fisher's exact test. Compares the distribution of epistemic and root *must* in each embedded environment to *must*'s distribution by flavor in a sample of 400 matrix declaratives (italicized). 'Complements of attitudes' comprises a random sample of 400 verb complements and all 134 adjective complements (italicized).