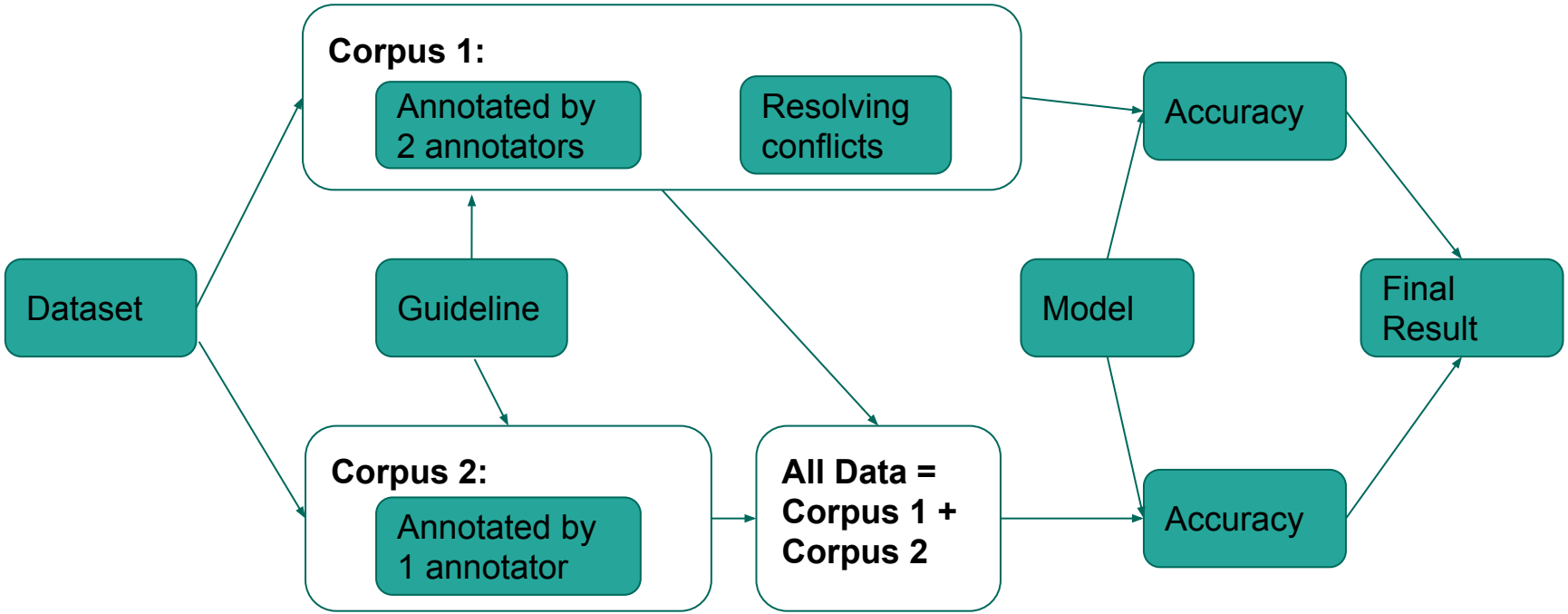
# Temporal Relations Annotation

Zhuoran Huang, Xiaoyu Lu, Yonglin Wang

# Goal of Project

- Find out the temporal relations between events in one paragraph
  - Events: each event is represented by a non-auxiliary verb in the sentence.
  - Temporal relations: before / after / simultaneously / vague
- Deliver a Gold Standard for annotating such temporal relations

# Pipeline


Corpus 1:
Annotated by 2 annotators
Resolving conflicts
Accuracy
Dataset
Guideline
Model
Final Result
Corpus 2:
Annotated by 1 annotator
All Data = Corpus 1 + Corpus 2
Accuracy

# Guideline

- **Relation Types:**
  For an event pair (A, B), choose, **based on the startpoints** of event A and event B, one of **the four TempRel types**:
  **BEFORE:** *E.g.The first thing I **ask** is that they **help** writing this column.*
  **AFTER:** *E.g.The first thing I **dislike** is that they **stop** writing this column.*
  **SIMULTANEOUS:** *E.g. As soon as the environment **becomes** stable again, the habit **starts** to reassert itself.*
  **VAGUE:** *E.g. Perhaps Mr. Sanders will **win** a party-run primary or caucus — in Alaska, Hawaii or Wyoming over the coming weeks — as he did in North Dakota last week. Otherwise Mr. Biden is the favorite in every remaining primary state. And without victories, Mr. Sanders will be **deprived** of opportunities to claim the momentum and favorable news coverage to change the trajectory of the race.*
- **Useful Diagnostics:**
  - **Generics:** If one word in a pair is generic while the other is non-generic, we should immediately label the pair as VAGUE.
  - **Possible World Analysis:** Each event should be treated as having occurred, whether or not the text implied that it had occurred. Negated events and hypotheticals are treated similarly. One assumes the event does occur, and all other events are ordered accordingly. Use context (word meaning or tense) to infer when the event would happen, if possible.
  - **Causality:** If event A causes/enables/prevents B, then we say A happens before B
  - **Condition:** If A is the necessary condition for B, then we say A happens before B

# Annotation Conflicts

- 275 out of 662 items have different annotations
- How to resolve conflicts:
  - Add one more blind annotation (team member annotate it without knowing others' annotation)
  - Choose the option with most votes
  - If there is still conflicts, team members discuss about it and make final decision.

| | **before** | **after** | **simultaneous** | **vague** | **total** |
|---|---|---|---|---|---|
| **before** | 103 | 6 | 11 | 51 | 171 |
| **after** | 23 | 41 | 8 | 39 | 111 |
| **simultaneous** | 7 | 3 | 10 | 11 | 31 |
| **vague** | 58 | 34 | 24 | 233 | 349 |
| **total** | 191 | 84 | 53 | 334 | 662 |

Inter-Annotator Agreement:

- Two annotations for one data point, total counts for each category are different: Chose Cohen's $\kappa$
- Cohen's $\kappa \approx 0.345$
- Almost in the fair/good range (.4-.75) according to Green (1997)
- Understandably low: lots of subjective judgements involved

# Model

- Features
  - Lexical Features
    - The part-of-speech (POS) tags from each individual verb and from its neighboring two words.
    - The modal verbs between the event mention (i.e., will, would, can, could, may and might).
    - The temporal connectives between the event mentions (e.g., before, after and since).
  - Syntactic Features
    - The distance between them in terms of the number of tokens.
  - Semantic Features
    - Whether the two verbs have a common synonym from their synsets in WordNet (Fellbaum, 1998).
    - Whether the input event mentions have a common derivational form derived from WordNet.
- Algorithm
  - Random forest

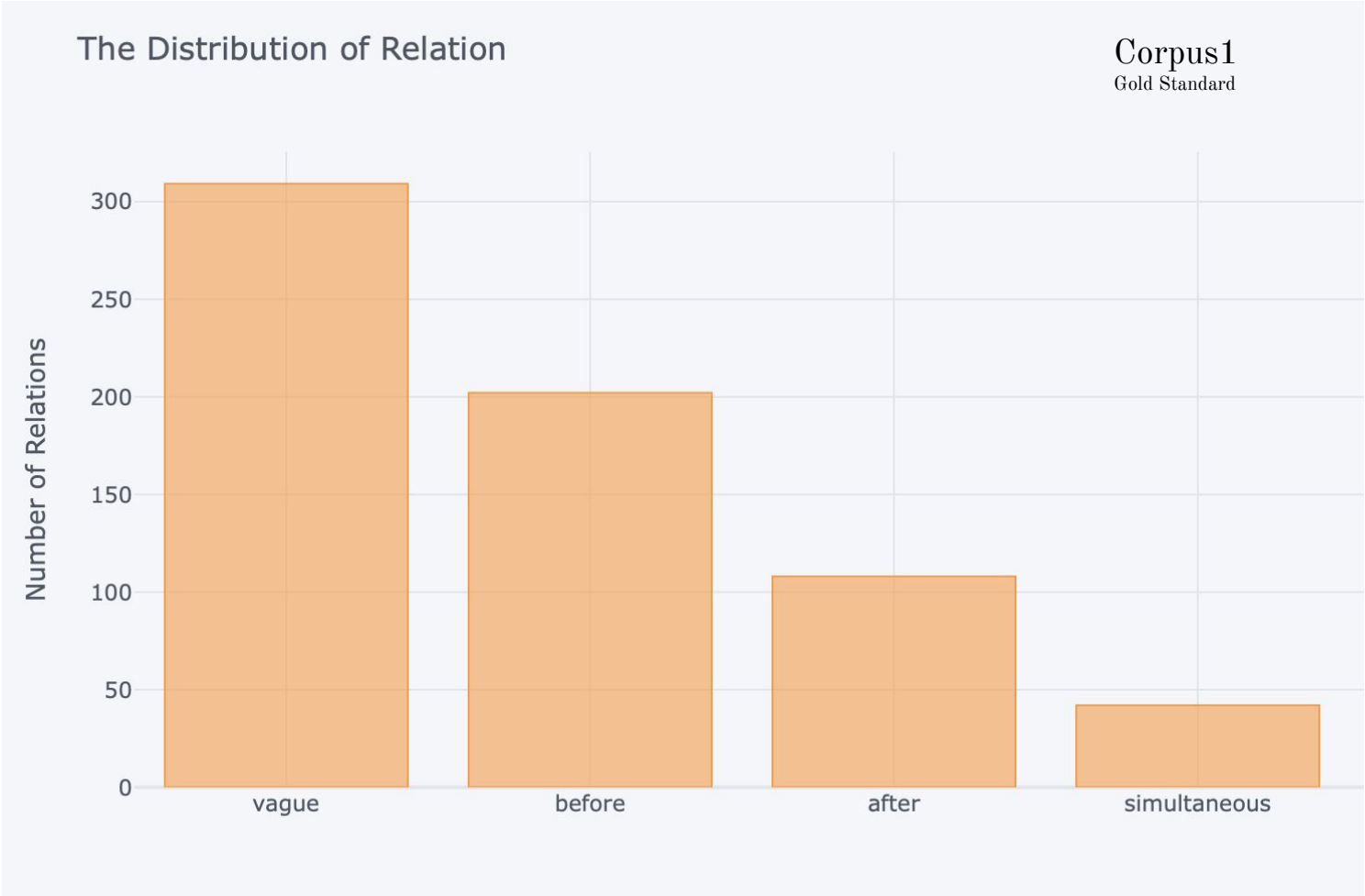
# Result Analysis

| Corpus 1 | precision | recall | f1 | support |
|---|---|---|---|---|
| vague | 0.82 | 0.54 | 0.65 | 94 |
| before | 0.40 | 0.55 | 0.46 | 31 |
| after | 0.11 | 0.40 | 0.17 | 5 |
| simu | 0 | 0 | 0 | 3 |

| ALL DATA | precision | recall | f1 | support |
|---|---|---|---|---|
| vague | 0.81 | 0.50 | 0.62 | 719 |
| before | 0.18 | 0.33 | 0.23 | 120 |
| after | 0.06 | 0.33 | 0.11 | 24 |
| simu | 0.12 | 0.42 | 0.19 | 31 |

662 vs 4467

- The overall accuracy of model in Corpus 1 is 0.53, in all data is 0.47
- Reason of low accuracy:
  - Imbalanced data distribution (too much “vague”)
    - Event extraction: there are some auxiliary verbs in it
    - Most events have no relations (hence labeled “vague”)
  - Inconsistent Annotation -> patterns not clear for model to learn
    - Guideline is not thorough
    - Temporal relations is quite ambiguous
  - Feature selection: can’t capture the essential features of temporal relations


The Distribution of Relation
Corpus1
Gold Standard
Number of Relations
300
250
200
150
100
50
0
vague
before
after
simultaneous

# Future Improvements

- Data
  - Fully delete auxiliary words
  - Annotate more data in the category of “before” “after” and “simultaneously”
- Annotation
  - Hire more annotators to annotate more data and each article should be annotated by more annotators
  - Include more ambiguous examples in the guideline
- Features
  - Add word vector features to the model

# Dataset

- Dataset: 7 NYT articles
- Data-processing:
  - Extracted event pairs for each paragraph
  - each event is represented by a non-auxiliary verb in the sentence.
  - *E.g. "as soon as the environment **becomes** stable again, the habit **starts** to **reassert** itself",*

| Index | Event 1 | Event 2 | Relation |
|---|---|---|---|
| | becomes | starts | |
| | becomes | reassert | |
| | starts | reassert | |

# THANK YOU

To our annotators for the hard work
And to everyone for listening

# Example of difficult judgement (from adjudication)

Text: Mr. Gergen suggested that some chief executives might want to lower their sights. “I think business chief executives tend to make better governors or mayors,” he **said**. “They can be more pragmatic and less ideological. Some of the names you’re hearing, I **have** trouble picturing them with the Elizabeth Warren or Bernie Sanders wing of the party. I wouldn’t call that a match made in heaven.”

**Judgement process:**

- Both are non-generic events --> comparable
- Which one happens first? Saying so or having trouble? Or at the same time? Or would more than one relations be acceptable to the best of your knowledge?
- Final adjudication: vague