# CORPUS CREATION FOR NEW GENRES:

## A Crowdsourced Approach to PP Attachment

COLUMBIA

Mukund Jha, Jacob Andreas, Kapil Thadani,
Sara Rosenthal, Kathleen McKeown

# Background



- Supervised techniques for text analysis require annotated data
- LDC provides annotated data for many tasks
- But performance degrades when these systems are applied to data from a different domain or genre

# This talk



> Can linguistic annotation tasks be extended to new genres at low cost?

# This talk



> Can **PP attachment annotation** be extended to **noisy web data** at low cost?

# Outline



1. Prior work
   - PP attachment
   - Crowdsourced annotation
2. Semi-automated approach
   - System: sentences → questions
   - MTurk: questions → attachments
3. Experimental study
4. Conclusion + Potential directions

# Outline



1. Prior work
   - PP attachment
   - Crowdsourced annotation
2. Semi-automated approach
   - System: sentences → questions
   - MTurk: questions → attachments
3. Experimental study
4. Conclusion + Potential directions

# PP attachment



- We went to John's house on Saturday

- We went to John's house on 12th street

- I saw the man with the telescope

# PP attachment



- So here my dears, is my top ten albums I heard in 2008 with videos and everything ( happily, the majority of these were in fact released in 2008, phew.)

# PP attachment



- PP attachment training typically done on RRR dataset (Ratnaparkhi et al., 1994)
  - Presumes the presence of an oracle to extract two potential attachments
  - eg: "cooked fish for dinner"
- PP attachment errors aren't well reflected in parsing accuracy (Yeh and Vilain, 1998)
- Recent work on PP attachment achieved 83% accuracy on the WSJ (Agirre et al., 2008)

# Crowdsourced annotations



- Can linguistic tasks be performed by untrained MTurk workers at low cost? (Snow et al., 2008) et al.
- Can PP attachment annotation be performed by untrained MTurk workers at low cost? (Rosenthal et al., 2010)
- Can PP attachment annotation be extended to noisy web data at low cost?

# Outline



1. Prior work
   - PP attachment
   - Crowdsourced annotation
2. Semi-automated approach
   - System: sentences → questions
   - MTurk: questions → attachments
3. Experimental study
4. Conclusion + Potential directions

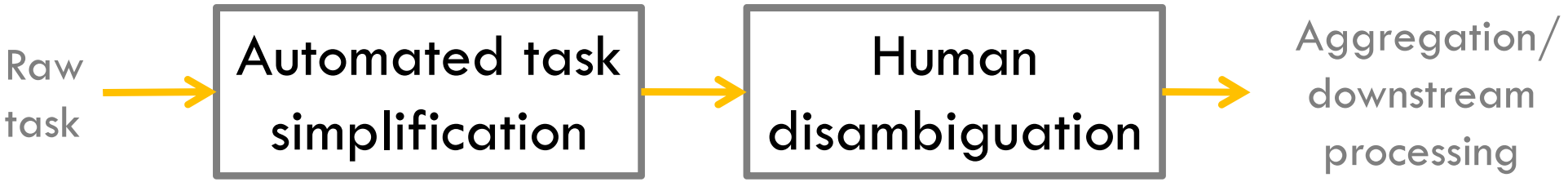
# Semi-automated approach



- Automated system
  - Reduce PP attachment disambiguation task to multiple-choice questions
  - Tuned for recall
- Human system (MTurk workers)
  - Choose between alternative attachment points
  - Precision through worker agreement

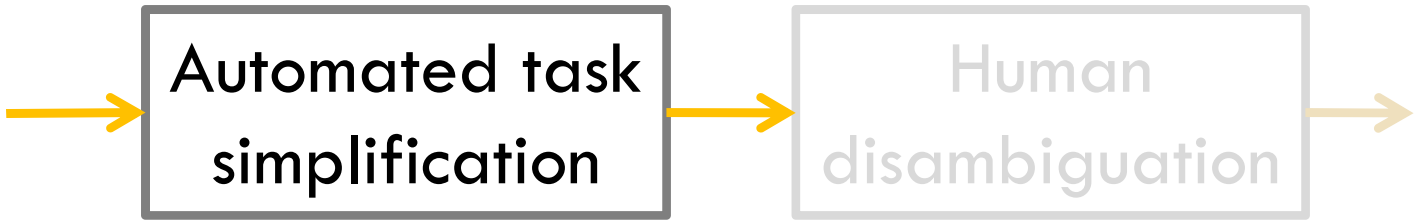
# Semi-automated approach





# Semi-automated approach





# Problem generation



1. Preprocessor + Tokenizer

2. CRF-based chunker (Phan, 2006)
   - Relatively domain-independent
   - Fairly robust to noisy web data

3. Identification of PPs
   - Usually Prep + NP
   - Compound PPs broken down into multiple simple PPs
   - eg: I just made some changes to the latest issue of our newsletter

# Attachment point prediction



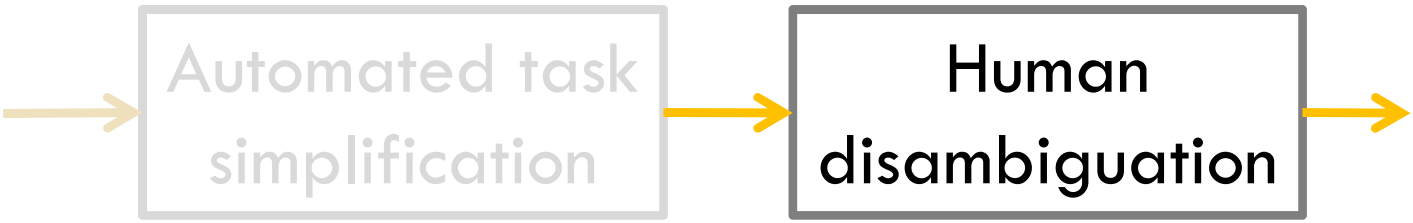
4. Identify potential attachment points for each PP
   - Preserve 4 most likely answers (give or take)
   - Heuristic-based

| | Rule | Example |
|---|---|---|
| 1. | Closest NP and VP preceding the PP | I made modifications **to our newsletter** |
| 2. | Preceding VP if closest VP contains a VBG | He snatched the disk flying away **with one hand** |
| 3. | First VP following the PP | **On his desk** he has a photograph |

… etc

# Semi-automated approach





# Mechanical Turk



**Instructions:**

Given below is a sentence with a prepositional phrase marked in red. Your task is to pick the phrase that is being modified by the given prepositional phrase.
(*Hovering over an answer will highlight it in the sentence*).
You are always required to choose an answer; however if you feel that the correct answer is not among the options or that the prepositional phrase is not well constructed, please let us know using the link below the options.

Show Examples

If that sort **of thing bores** you , this post would be a good time to go out to the lobby and get yourself a snack .

**Consider the sentence above. Which of the following does the prepositional phrase of thing bores modify?**

- ○ would be
- ○ to go out
- ◉ that sort

Click here to hide these options.

**Tick the following options regarding the question:**
(Note: You are still required to pick the best option from the choices above)

- ☐ Correct answer is not present in the above choices
- ☑ Prepositional phrase is not correct

Enter the correct answer:

Enter the correct prepositional phrase: of thing

Please provide any comments you may have below, we appreciate your input!

submit

# Mechanical Turk



**Instructions:**

Given below is a sentence with a prepositional phrase marked in red. Your task is to pick the phrase that is being modified by the given prepositional phrase. (*Hovering over an answer will highlight it in the sentence*).
You are always required to choose an answer; however if you feel that the correct answer is not among the options or that the prepositional phrase is not well constructed, please let us know using the link below the options.

Show Examples

If that sort **of thing bores** you , this post would be a good time to go out to the lobby and get yourself a snack .

**Consider the sentence above. Which of the following does the prepositional phrase of thing bores modify?**

- ○ would be
- ○ to go out
- ◉ that sort

Click here to hide these options.

**Tick the following options regarding the question:**
(Note: You are still required to pick the best option from the choices above)

- ☐ Correct answer is not present in the above choices
- ☑ Prepositional phrase is not correct

Enter the correct answer:
Enter the correct prepositional phrase: of thing

Please provide any comments you may have below, we appreciate your input!

submit

# Outline



1. Prior work
   - PP attachment
   - Crowdsourced annotation
2. Semi-automated approach
   - System: sentences → questions
   - MTurk: questions → attachments
3. Experimental study
4. Conclusion + Potential directions

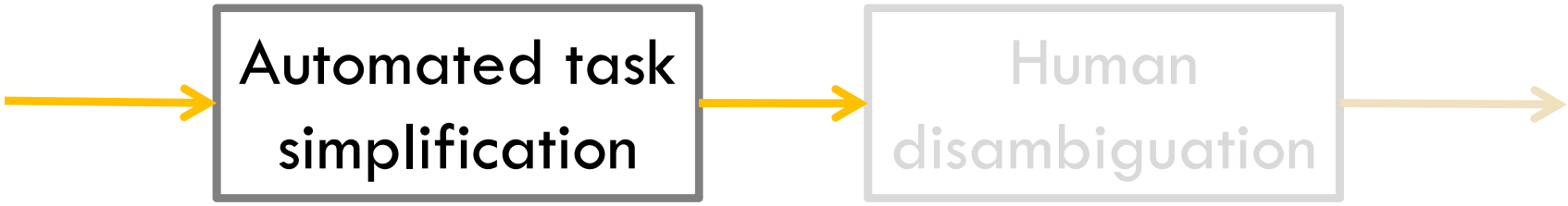
# Experimental setup



- Dataset: LiveJournal blog posts
- 941 PP attachment questions
- Gold PP annotations:
  - Two trained annotators
  - Disagreements resolved by annotator pool
- MTurk study:
  - 5 workers per question
  - Avg time per task: 48 seconds

# Results: Attachment point prediction

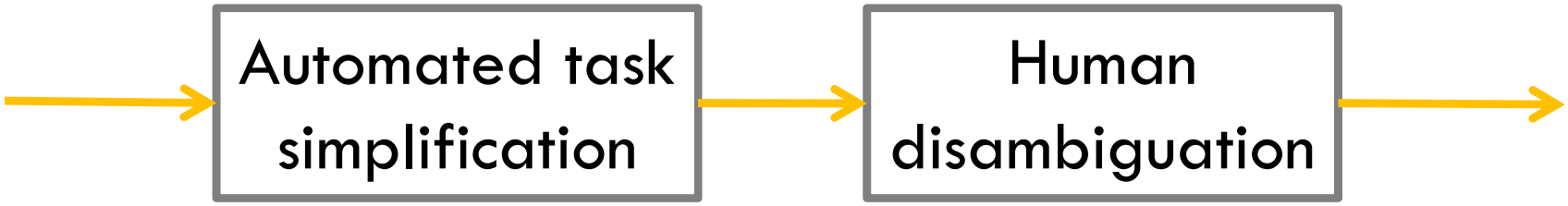





- Correct answer among options in 95.8% of cases
  - 35% of missed answers due to chunker error
  - But in 87% of missed answer cases, at least one worker wrote in the correct answer

# Results: Full system

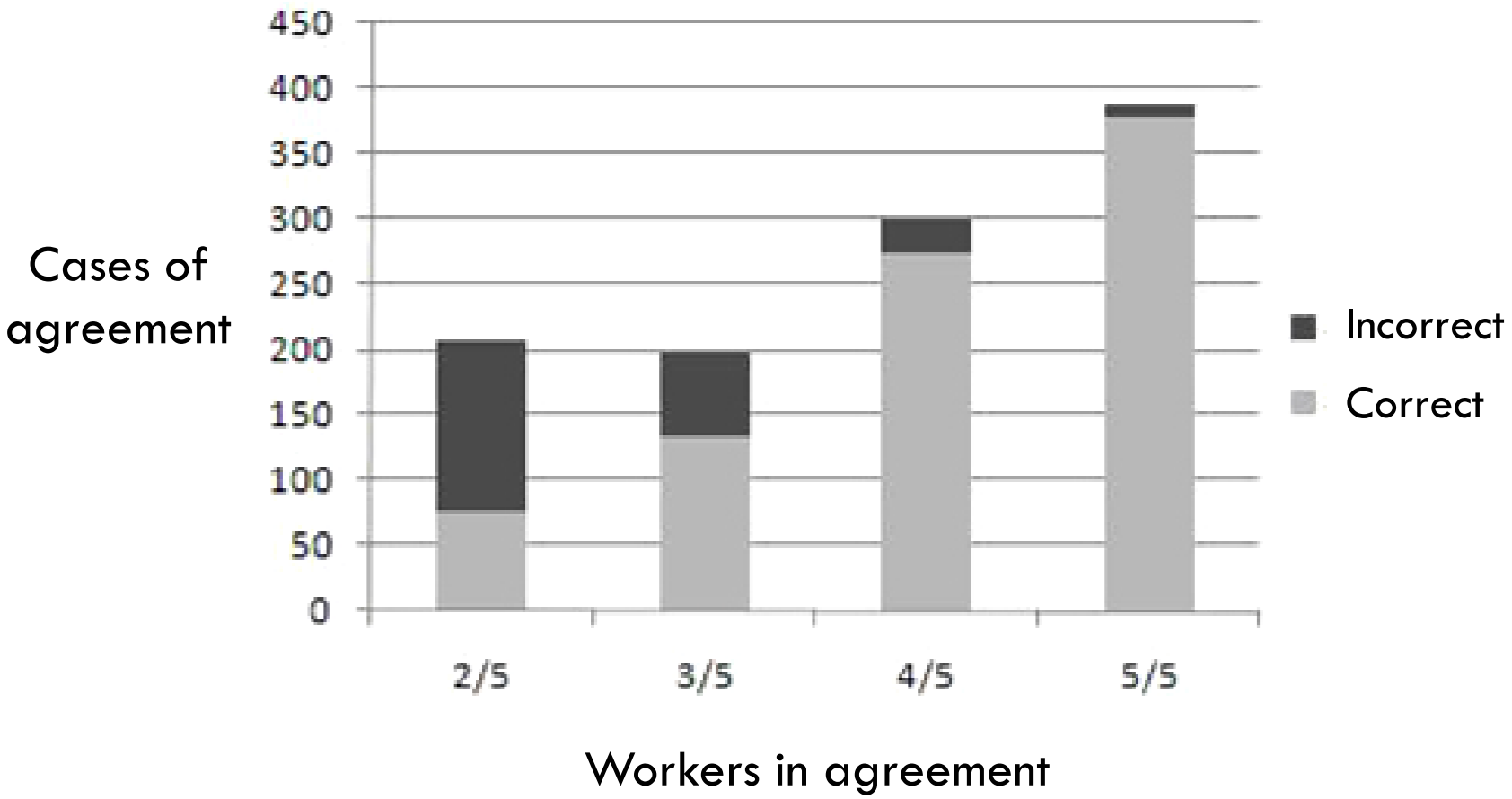





- Accurate attachments in 76.2% of all responses
  - Can we do better using inter-worker agreement?

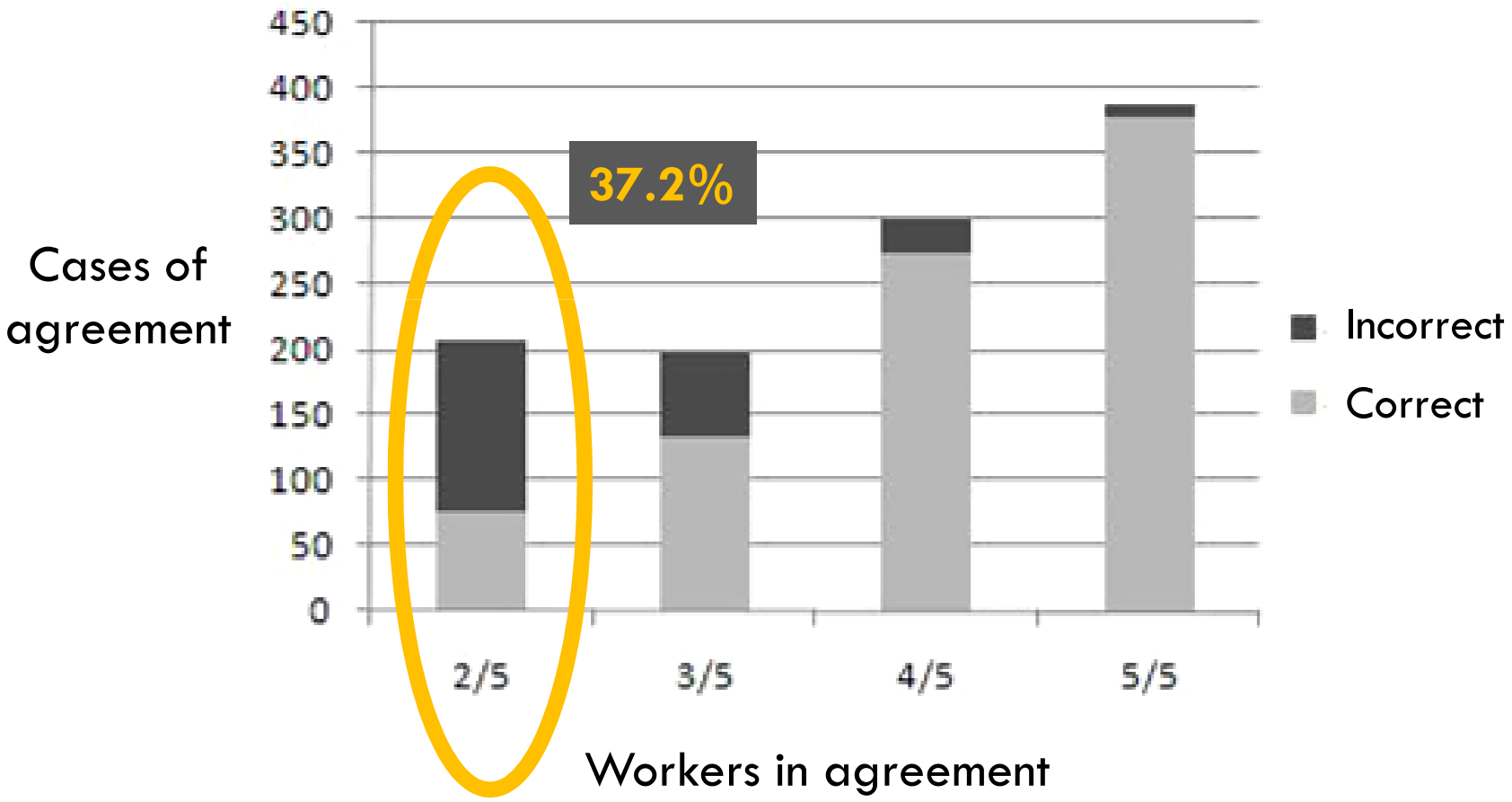
# Results: By agreement





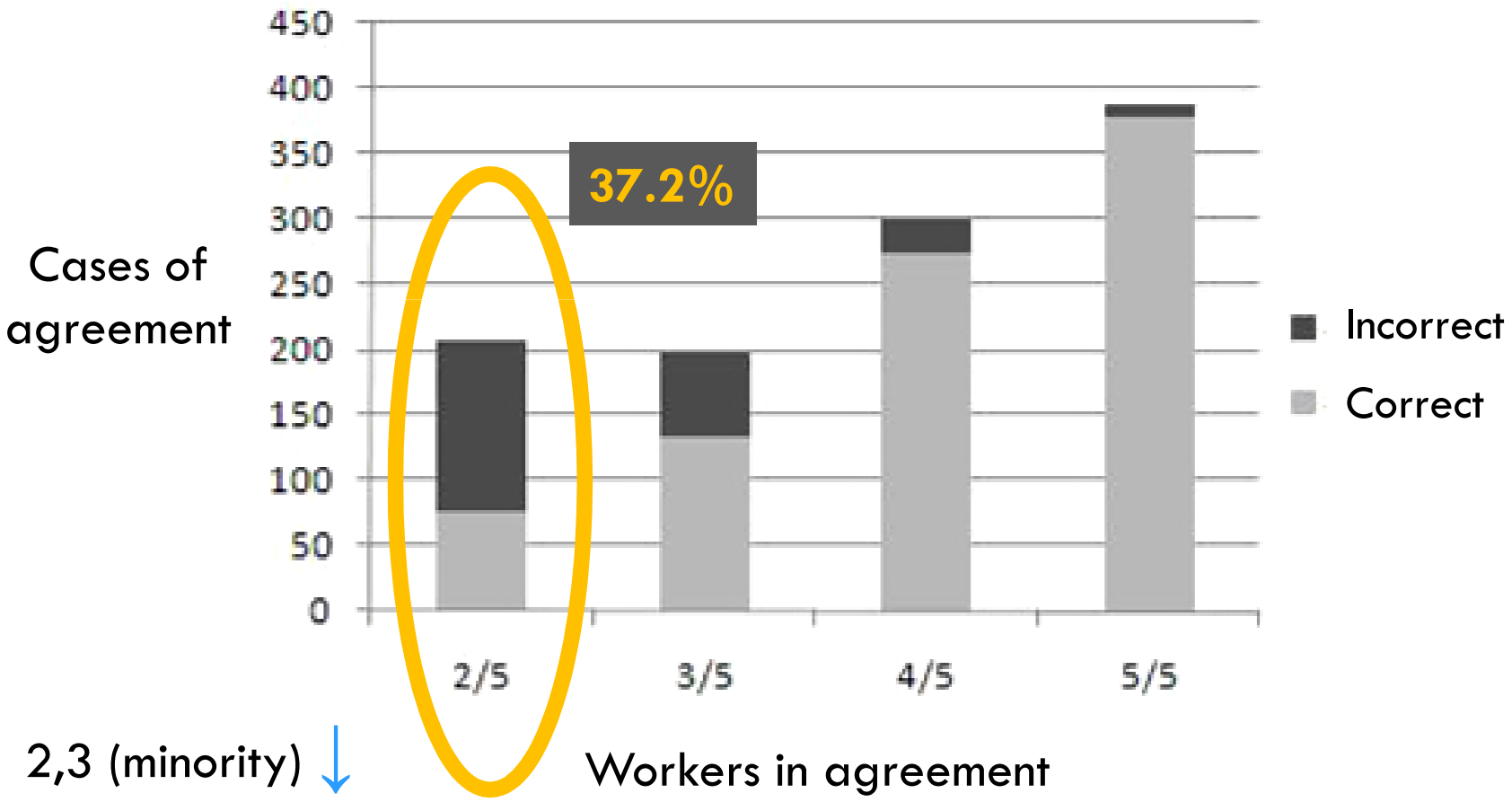
# Results: By agreement





# Results: By agreement






- 2,3 (minority) ↓
- 2,2,1 ↔
- 2,1,1,1 (plurality) ↑ **64.3%**

# Results: Cumulative



| Workers in agreement | Number of questions | Accuracy | Coverage |
|---|---|---|---|
| 5 | 389 | 0.97 | 41% |
| ≥ 4 | 689 | 0.95 | 73% |
| ≥ 3 | 887 | 0.89 | 94% |
| ≥ 2 (pl) | 906 | 0.88 | 96% |
| **All** | **941** | **0.84** | **100%** |
| (Rosenthal et al., 2010) | | 0.92 | |

# Results: Cumulative



| Workers in agreement | Number of questions | Accuracy | Coverage |
|---|---|---|---|
| 5 | 389 | 0.97 | 41% |
| ≥ 4 | 689 | 0.95 | 73% |
| ≥ 3 | 887 | 0.89 | 94% |
| ≥ 2 (pl) | 906 | 0.88 | 96% |
| **All** | **941** | **0.84** | **100%** |
| (Rosenthal et al., 2010) | | 0.92 | |

# Results: Cumulative



| Workers in agreement | Number of questions | Accuracy | Coverage |
|---|---|---|---|
| 5 | 389 | 0.97 | 41% |
| ≥ 4 | 689 | 0.95 | 73% |
| ≥ 3 | 887 | 0.89 | 94% |
| ≥ 2 (pl) | 906 | 0.88 | 96% |
| **All** | **941** | **0.84** | **100%** |
| (Rosenthal et al., 2010) | | 0.92 | |

# Results: Cumulative



| Workers in agreement | Number of questions | Accuracy | Coverage |
|---|---|---|---|
| 5 | 389 | 0.97 | 41% |
| ≥ 4 | 689 | 0.95 | 73% |
| ≥ 3 | 887 | 0.89 | 94% |
| ≥ 2 (pl) | 906 | 0.88 | 96% |
| **All** | **941** | **0.84** | **100%** |
| (Rosenthal et al., 2010) | | 0.92 | |

# Results: Cumulative

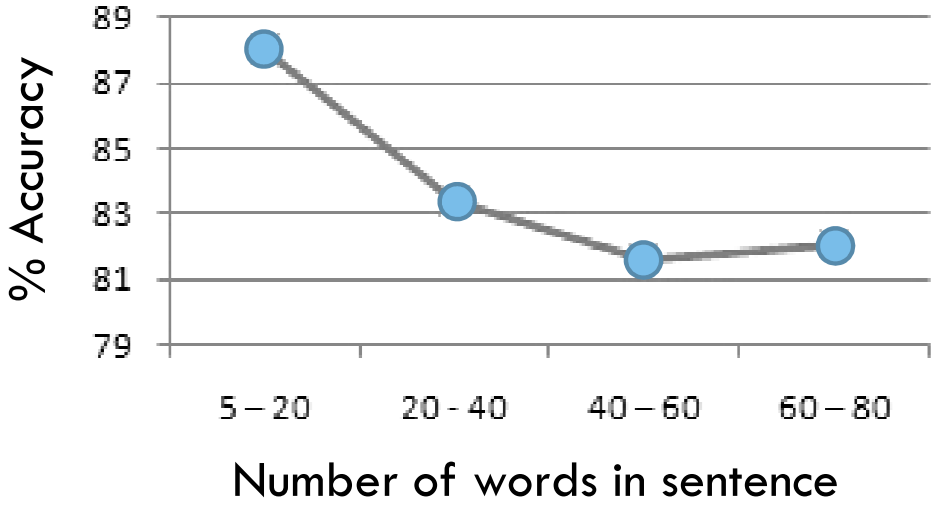


| Workers in agreement | Number of questions | Accuracy | Coverage |
|---|---|---|---|
| 5 | 389 | 0.97 | 41% |
| ≥ 4 | 689 | 0.95 | 73% |
| ≥ 3 | 887 | 0.89 | 94% |
| ≥ 2 (pl) | 906 | 0.88 | 96% |
| **All** | **941** | **0.84** | **100%** |
| (Rosenthal et al., 2010) | | 0.92 | |

# Results: Factors affecting accuracy



- Variation with length of sentence




- Variation with number of options

| No. of options | No. of cases | Accuracy |
|---|---|---|
| < 4 | 179 | 0.866 |
| 4 | 718 | 0.843 |
| > 4 | 44 | 0.796 |

# Outline

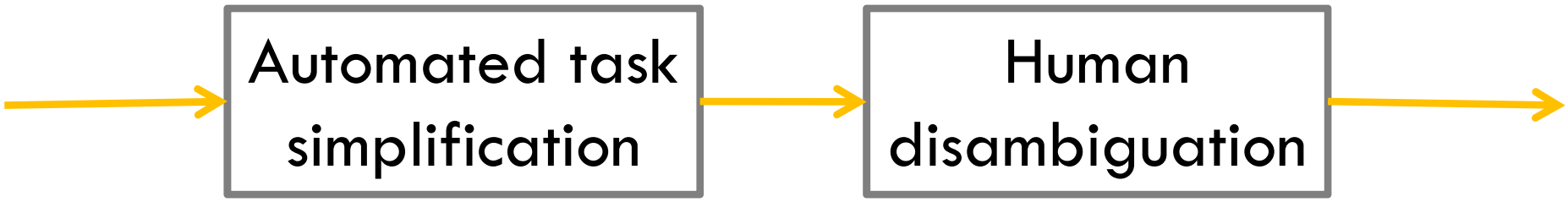


1. Prior work
   - PP attachment
   - Crowdsourced annotation
2. Semi-automated approach
   - System: sentences → questions
   - MTurk: questions → attachments
3. Experimental study
4. Conclusion + Potential directions

# Conclusion



- Constructed a corpus of PP attachments over noisy blog text
- Demonstrated a semi-automated mechanism for simplifying the human annotation task

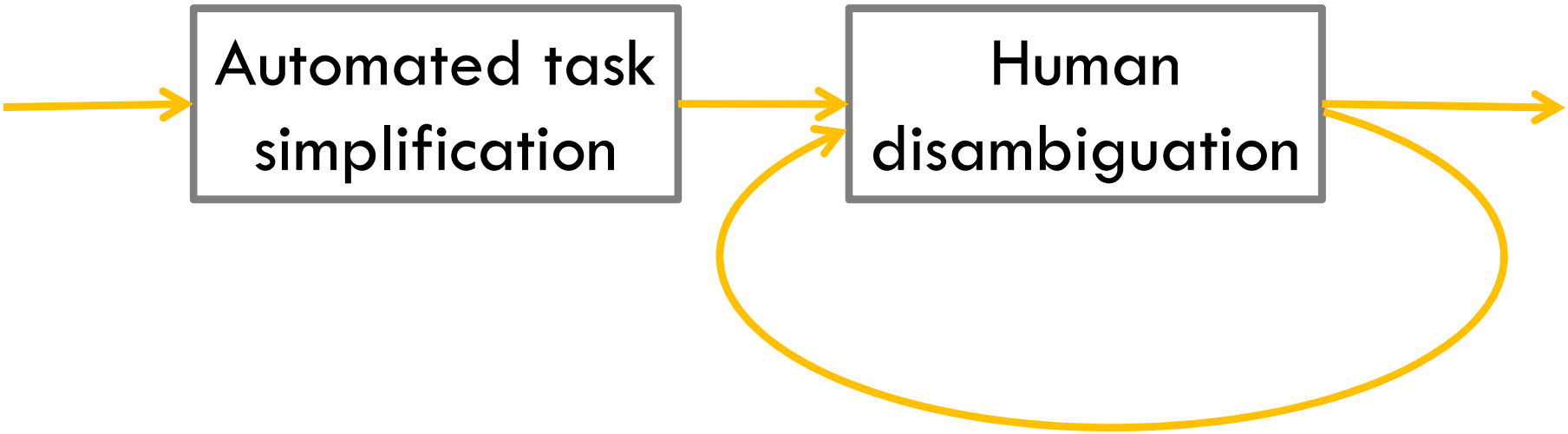



- Shown that MTurk workers can disambiguate PP attachment fairly reliably, even in informal genres

# Future work



- Use agreement information to determine when more judgements are needed

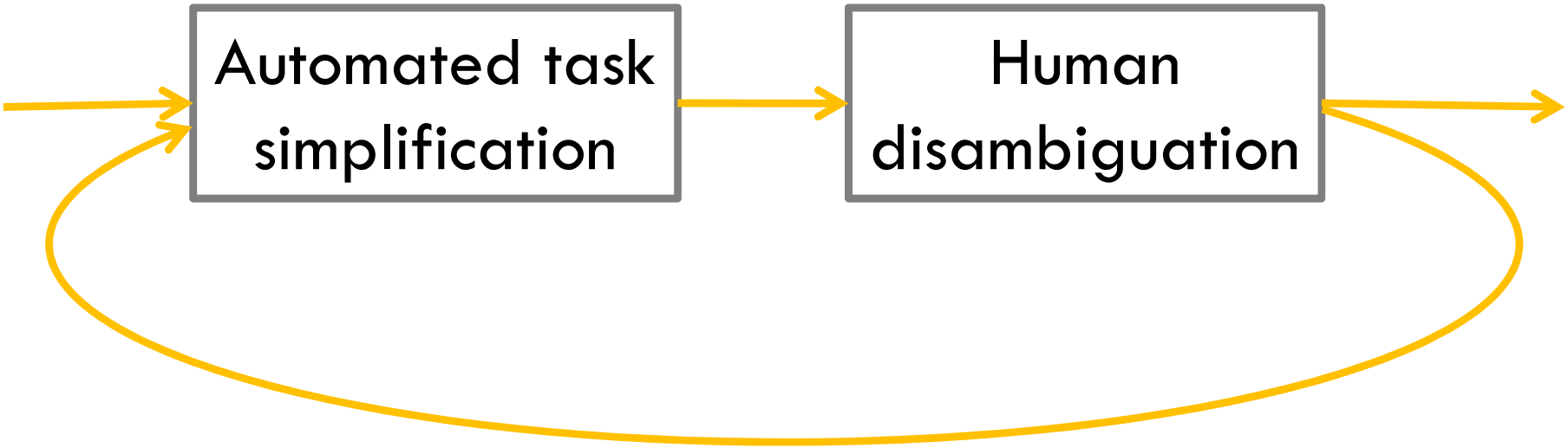



- Low agreement cases
- Expected harder cases (#words, #options)

# Future work



- Use worker decisions, corrections to update automated system




- Corrected PP boundaries
- Missed answers
- Statistics for attachment model learner

…

# Thanks