Yen-Ting Wang

Master Thesis Student

@ SAP TI AI PS FrontRunner - **Eduardo & Martin**

@ Mannheim University - **Prof. Han van der Aa & Ph.D. Adrian Rebmann**

Quick Refresher

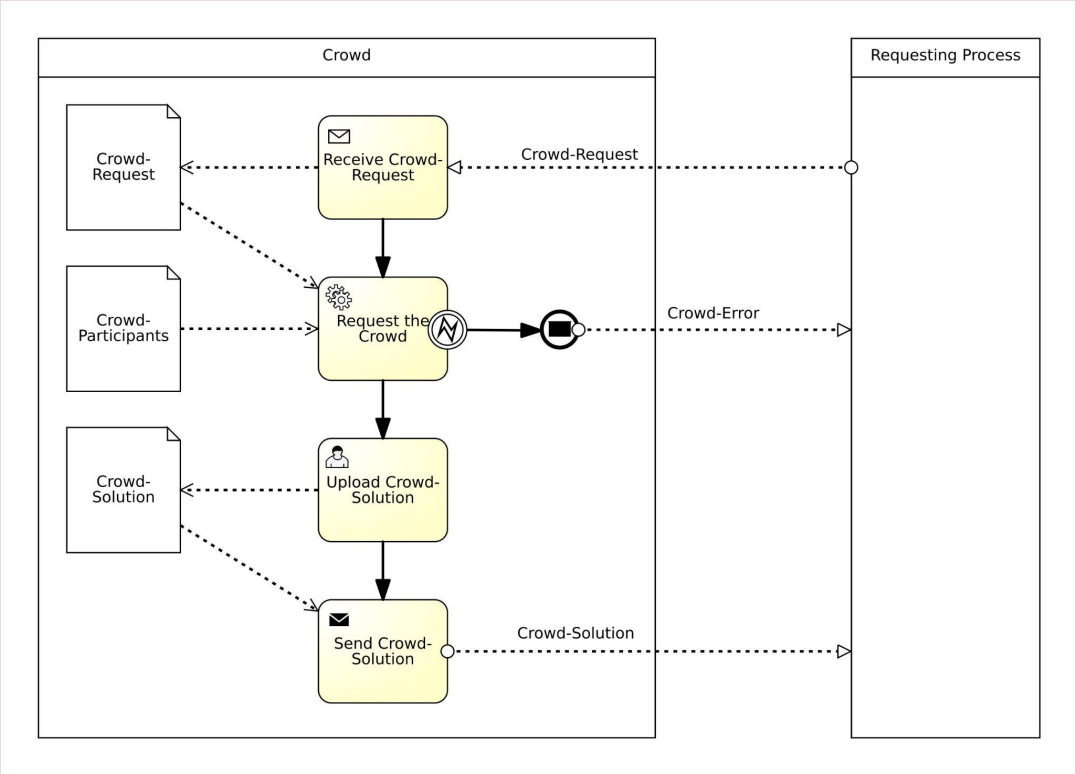
Automatic Generation of Activity Labels for Event Abstraction (Business Process Model Summarization)

0

Quick Refresher

- **Text Summarization** techniques for *High-level Activity Process Label*
  - **Pegasus**: Abstractive summarizer - important sentence masking as pre-training task
  - **Pegasus-TML**: Triplet Margin Loss as additional pre-training task
  - **Pegasus-Aug**: Data augmentation - round-trip translation applied
- **Auto Eval** using metrics such as **Rouge or BertScore**
  - Cannot distinguish performances
    - between **different models**
    - against **human generated labels**

Example 1



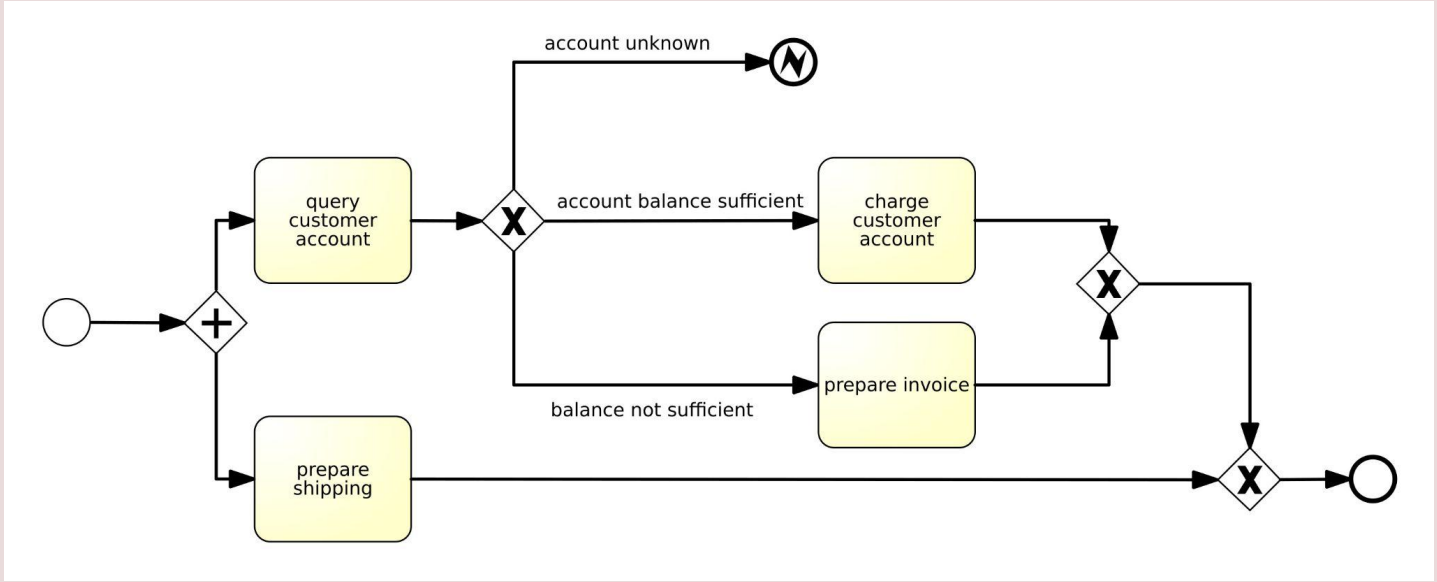
request crowd-based solutions

upload crowd-request

handle crowd request

operate problem solving process

Example 2



prepare shipping invoice

prepare customer order

manage customer account configuration

process shipping order request

The main takeaways from last presentation

- Labeling automatically for processes is hard
- Utilizing pre-trained language model achieves good results with limited data
- What we achieve now is being able to offer valid suggestions to users
  - View different perspectives of suggestions
  - Select or combine the suggestions
- Perhaps offer ranking for the future

Evaluation Survey

Received in total 31 surveys completed

- 24 fully completed (all 10 cases evaluated)
- 7 half completed (first 5 cases evaluated)

1

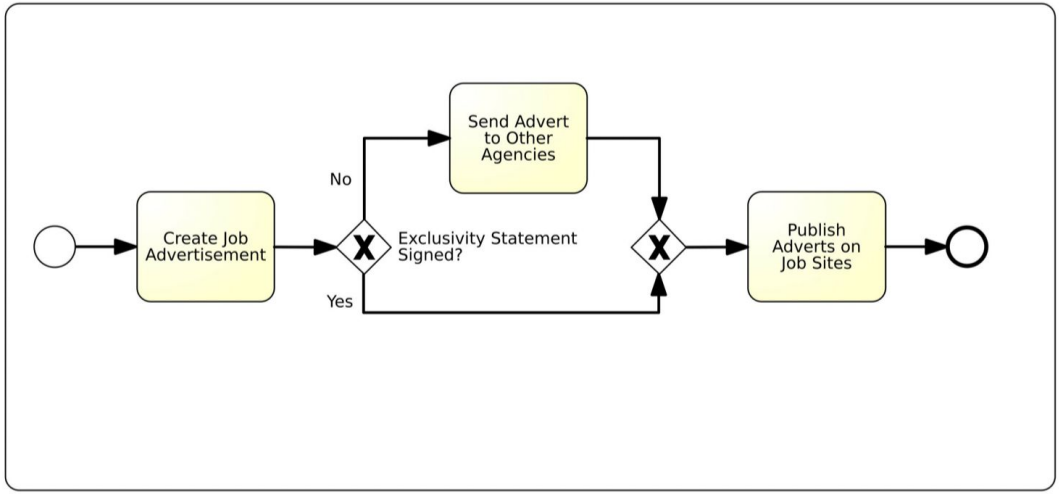
Survey Run-through

2

Case 1. Extracted labels

create job advertisement, send advert to other agencies, publish adverts on job sites

Tip: In our project setting, only the labels of events (circular elements) and tasks (rectangle elements) in the process model are extracted. Due to data quality issues, not all the labels of events and tasks can be extracted successfully.



*** 3. Choose the most suitable generated label for the process presented: (Select one option)**

- ○ create job advertisement
- ○ create and post advertisement
- ○ create and publish job advert
- ○ publish advertisement
- ○ none of the above is relevant (if selected, please skip all questions below and go to next page)

4. Main reason for the labelling choice (one or more is possible):

- ☐ Most important task
- ☐ Main outcome
- ☐ Provides a good overview

5. Rate the selected label:
Tip: Relevance - How applicable is the label to the process presented (Precision)? Informativeness - Is the label a sufficient representative of the process presented (Recall)? Fluency - Is the label easy to read and understand?

	Lowest 1	2	3	4	Highest 5
(a) Relevance	○	○	○	○	○
(b) Informativeness	○	○	○	○	○
(c) Fluency	○	○	○	○	○

Manually created label

publish job advertisement

*** 6. What do you think about the quality of the selected generated label in comparison to the manually created label? (Select one option)**

- ○ Better than the manually created label
- ○ Equally good
- ○ Worse

7. Rate the manually created label:
Tip: Relevance - How applicable is the label to the process presented (Precision)? Informativeness - Is the label a sufficient representative of the process presented (Recall)? Fluency - Is the label easy to read and understand?

	Lowest 1	2	3	4	Highest 5
(a) Relevance	○	○	○	○	○
(b) Informativeness	○	○	○	○	○
(c) Fluency	○	○	○	○	○

8. Do you have an alternative labeling suggestion for the model fragment? (optional)

Human Evaluation & Error Analysis

Report Outline

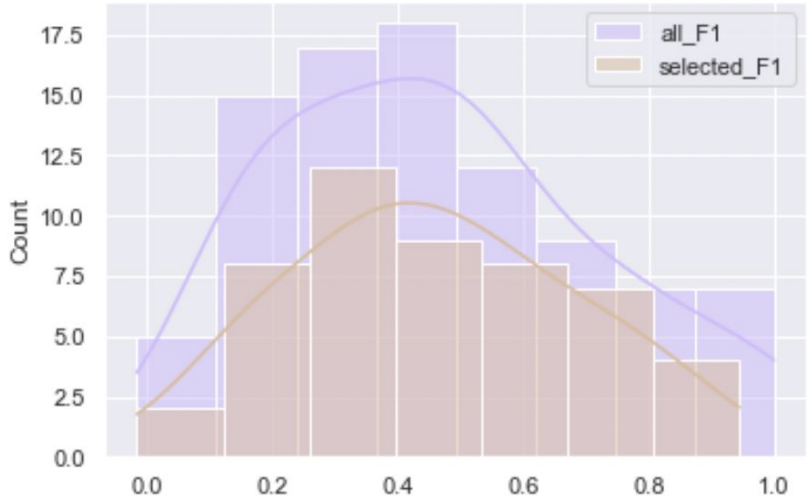
- **Case selection** for human evaluation
- **Analysis** on human evaluation results
  - **Performance comparison**
    - Model (M) vs. Human (H) generated labels
      - Direct comparison (better, equal, worse)
      - Metric scores comparison (relevance, informativeness, fluency)
    - Between different models
  - **Evaluators' main reasons** when selecting labels/models
  - **Error analysis**
    - Look into individual cases
  - **Inter-rater reliability**
    - Check agreement between evaluators

Case selection for human evaluation

0

Select
50 examples
out of 90
in test set

Based on test set distribution of BertScore F1 evaluation metric



Examples with full score are filtered out.
Full score means all model generated labels in an example are identical to its human label.

Performance comparison

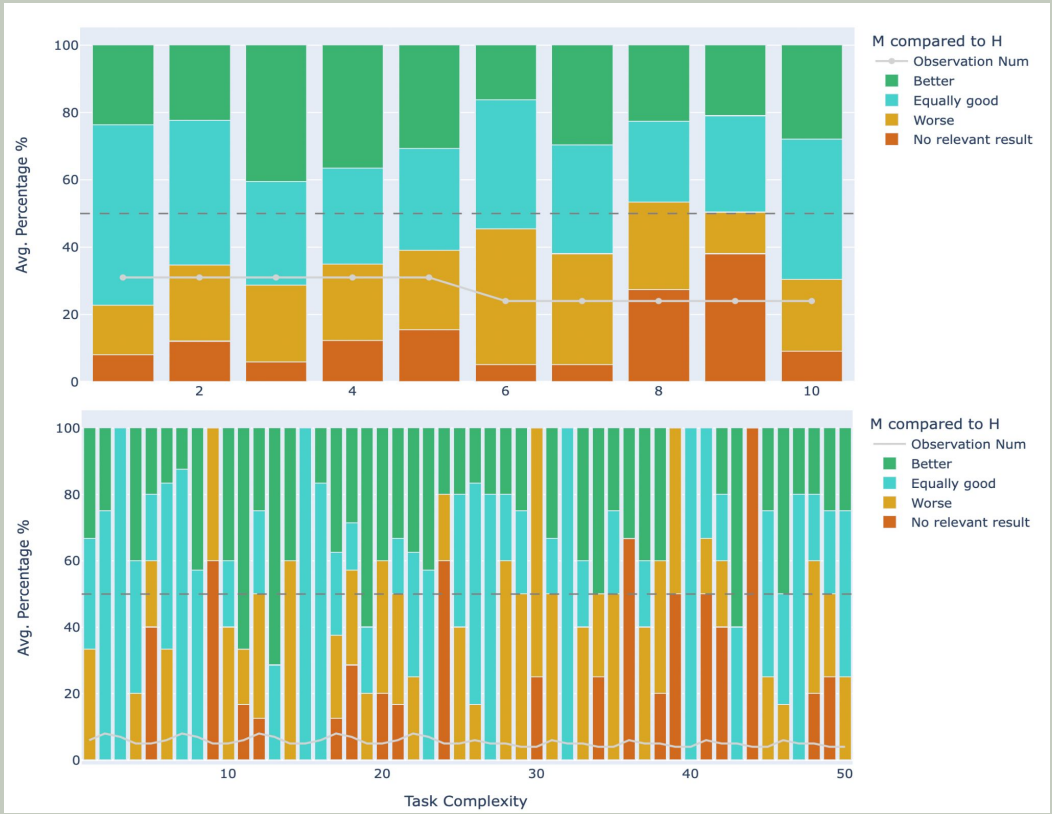
Direct comparison between Model (M) and Human (H) generated labels (M better than H, equally good, or worse)

1

Direct Comparison: M vs. H

After accounting for 10 contradictory answers out of total 275 (~3.6%)

(labeled worse but model generated labels contain one identical to human generated label)



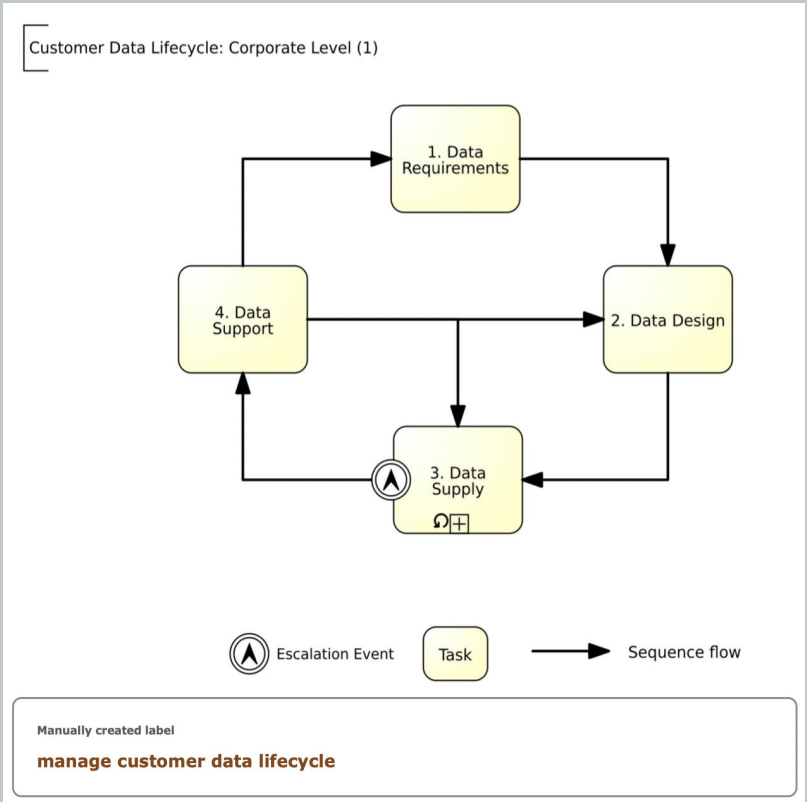
"No relevant result" could be due to data quality issue or the underlying process unclear to evaluators

Error Analysis

Look into individual cases labeled with worse performance

2

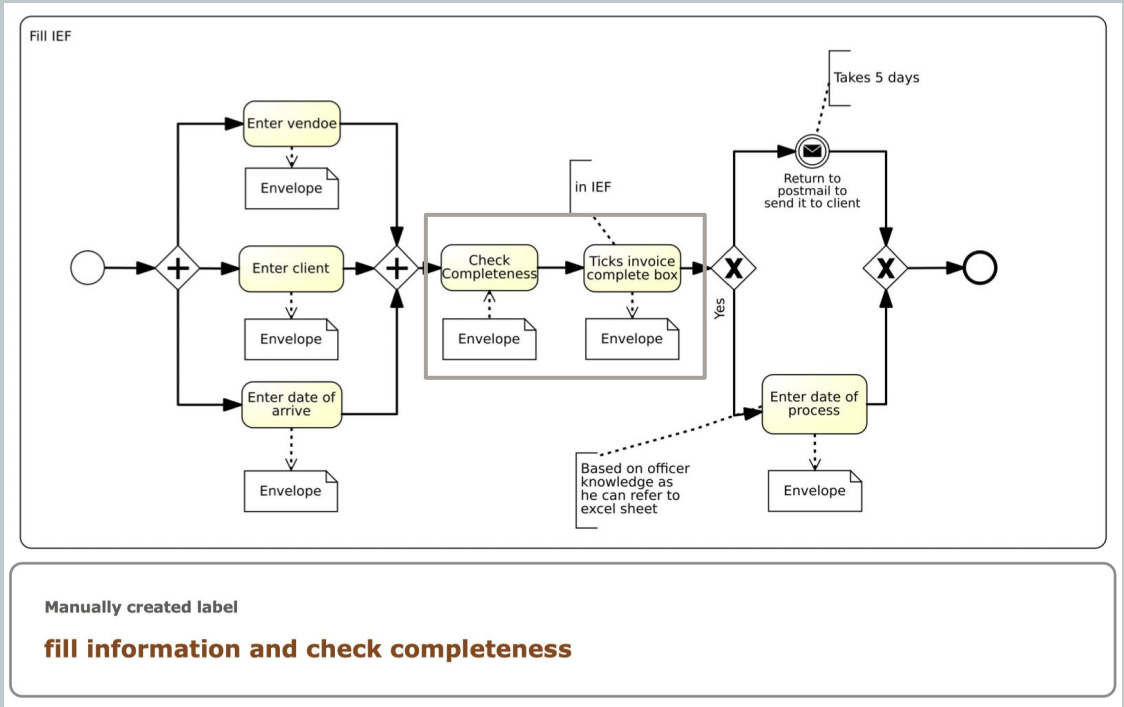
Task 9: 60% “no relevant result” & 40% “worse”



- manage data requirements
- perform data analysis to conceptualise data
- perform data analysis
- manage data supply

Feedback received: Underlying process could be unclear to evaluator(s)

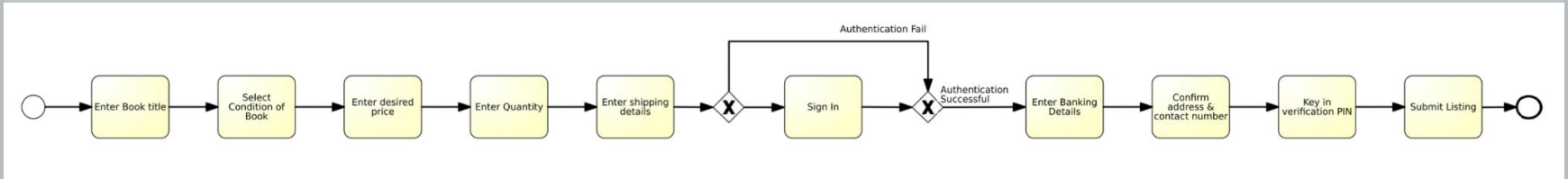
Task 30: 75% “worse” & 25% “no relevant result”



- complete invoice entry form
- tick the invoice box
- complete invoice configuration

Not all important tasks are covered by the model generated labels

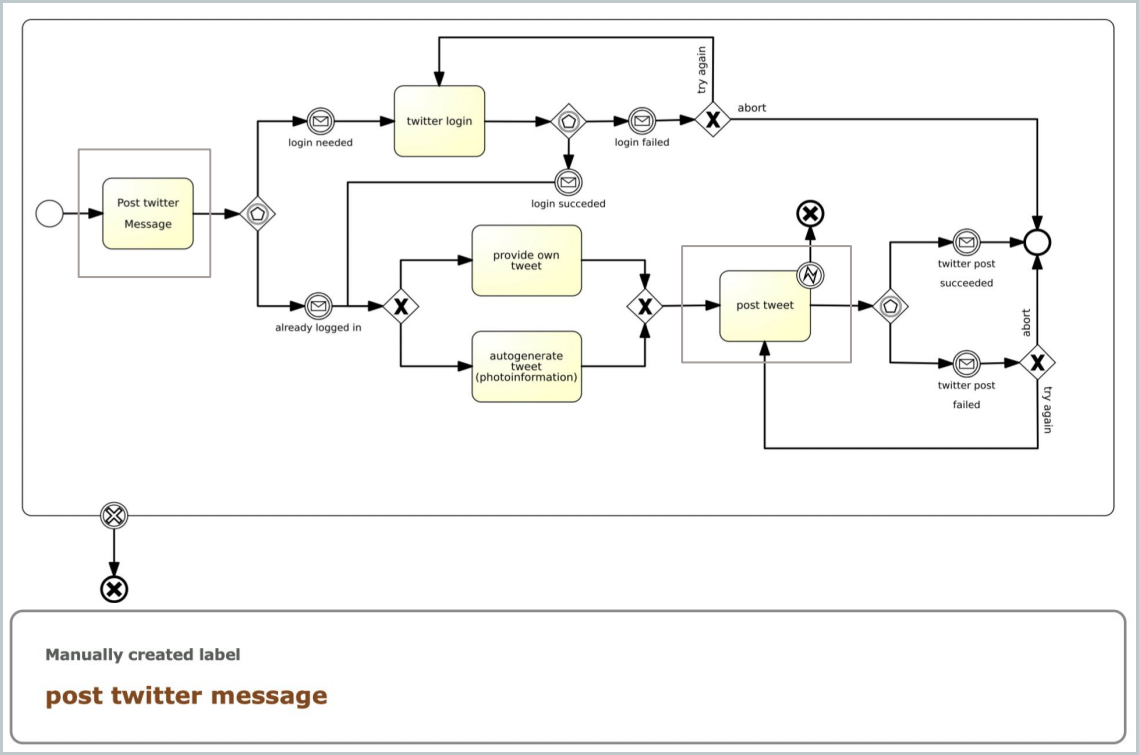
Task 39: 50% “worse” & 50% “no relevant result”



- enter book title
- apply for the first time
- registration for the book
- book seller ebay account

Models are not yet able to (or trained to) produce labels with high abstraction level
-> contextual information missing

Task 44: all evaluators vote for “no relevant result”



- log in
- handle twitter login
- login twitter

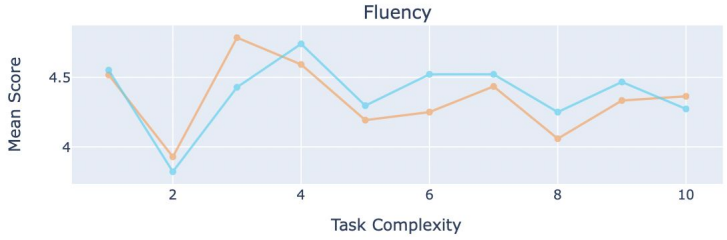
In the future: Less weight given to events and focus more on activities

Performance comparison

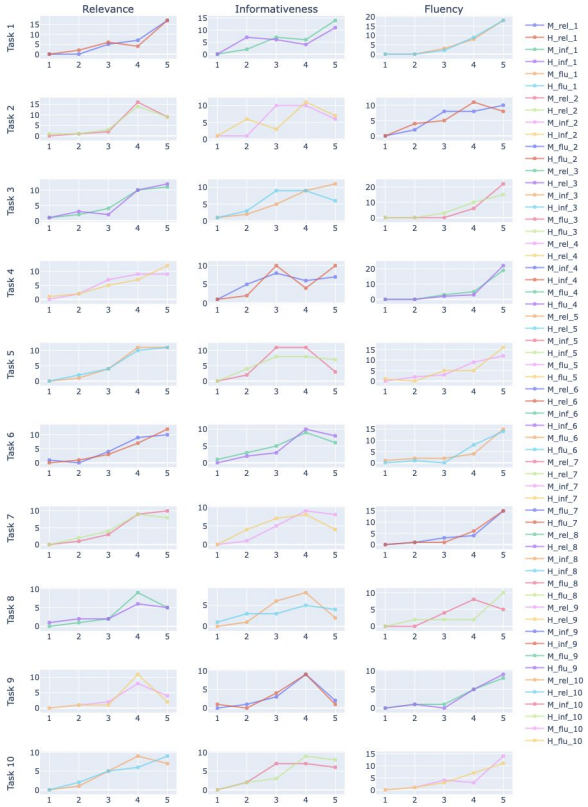
Metric Scores comparison between Model (M) and Human (H) generated labels (Relevance, Informativeness, Fluency)

3

Metric Scores Mean: M vs. H



Metric Scores Count: M vs. H



Metric Scores: M vs. H

	Relevance	Informativeness	Fluency
M_mean	**3.99**	**3.66**	4.29
H_mean	3.95	3.64	**4.35**
M_std	**0.75**	**0.82**	**0.63**
H_std	0.82	0.86	0.66

Inter-Rater Reliability

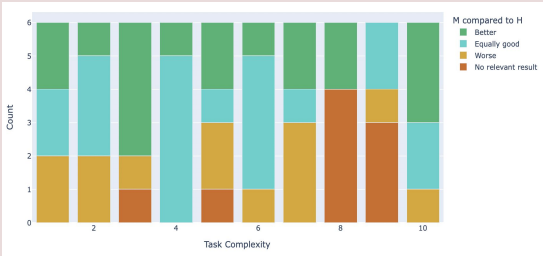
How much do evaluators agree with each other?

4

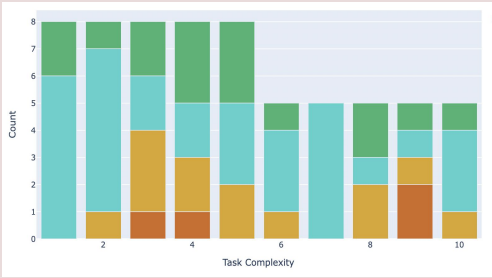
Distribution of Selected Labels sorted by Popularity

Count of Direct Comparison

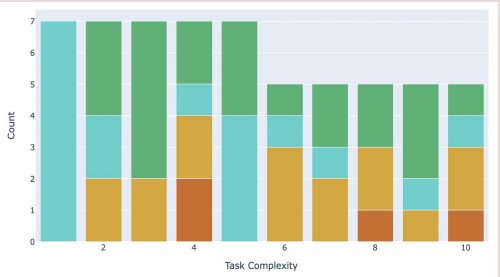
Batch 1



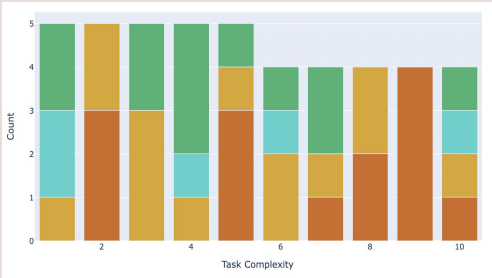
Batch 2



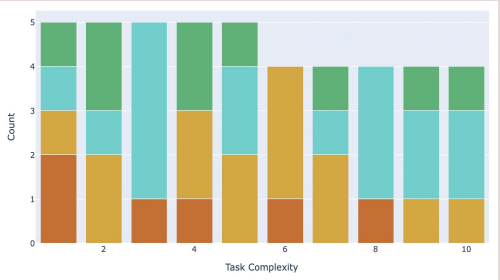
Batch 3



Batch 4

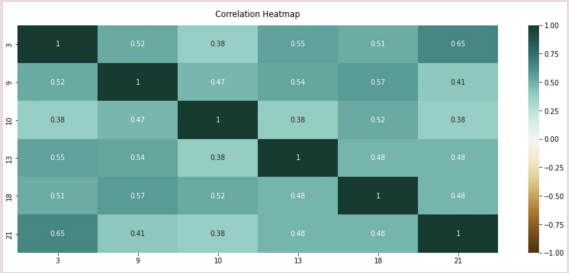


Batch 5

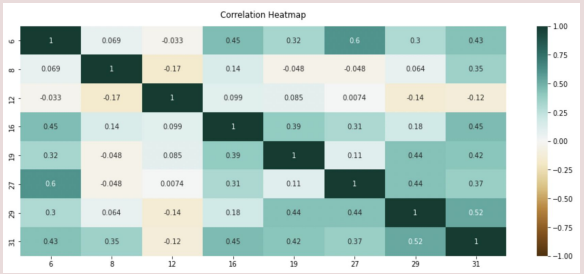


Pearson Correlation of Metric Scores

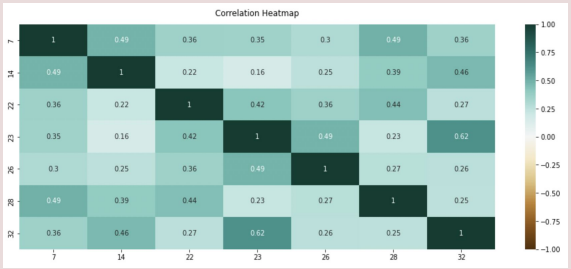
Batch 1



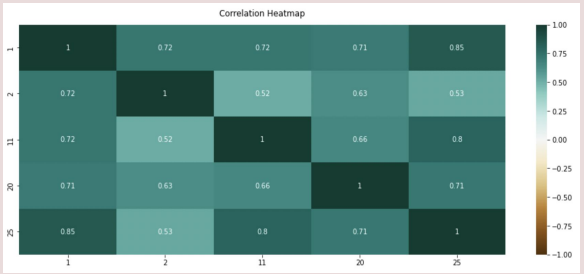
Batch 2



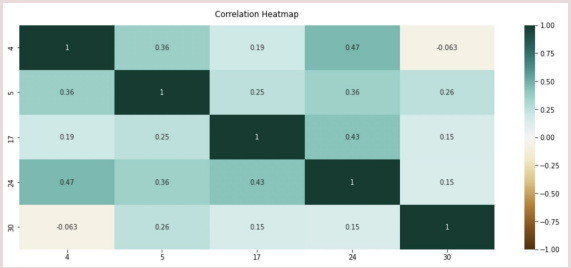
Batch 3



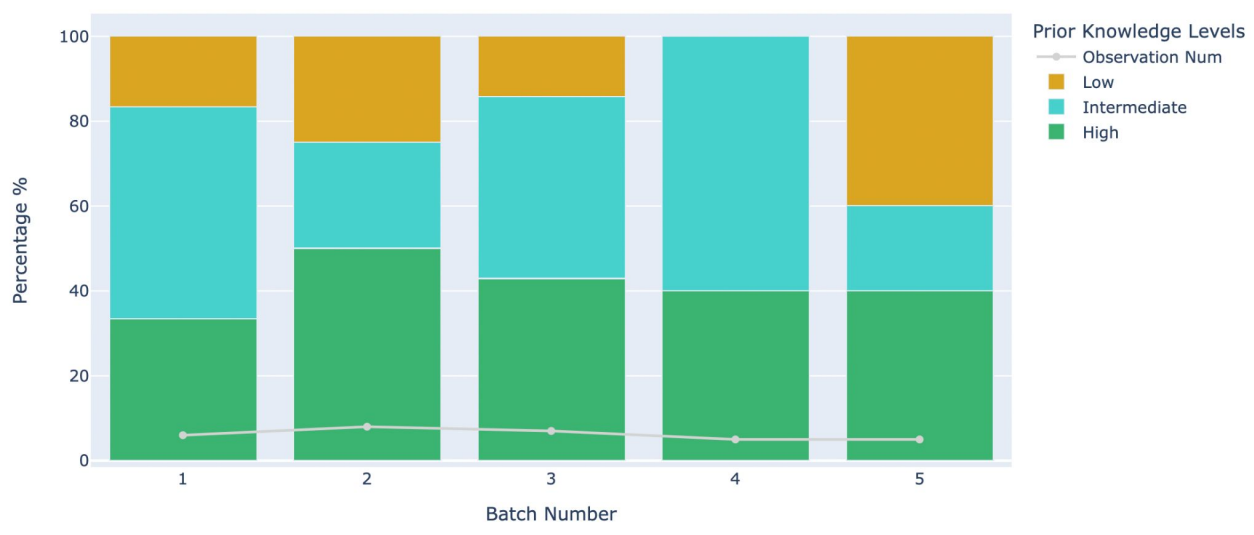
Batch 4



Batch 5



Distribution of Evaluators' Prior Knowledge Levels

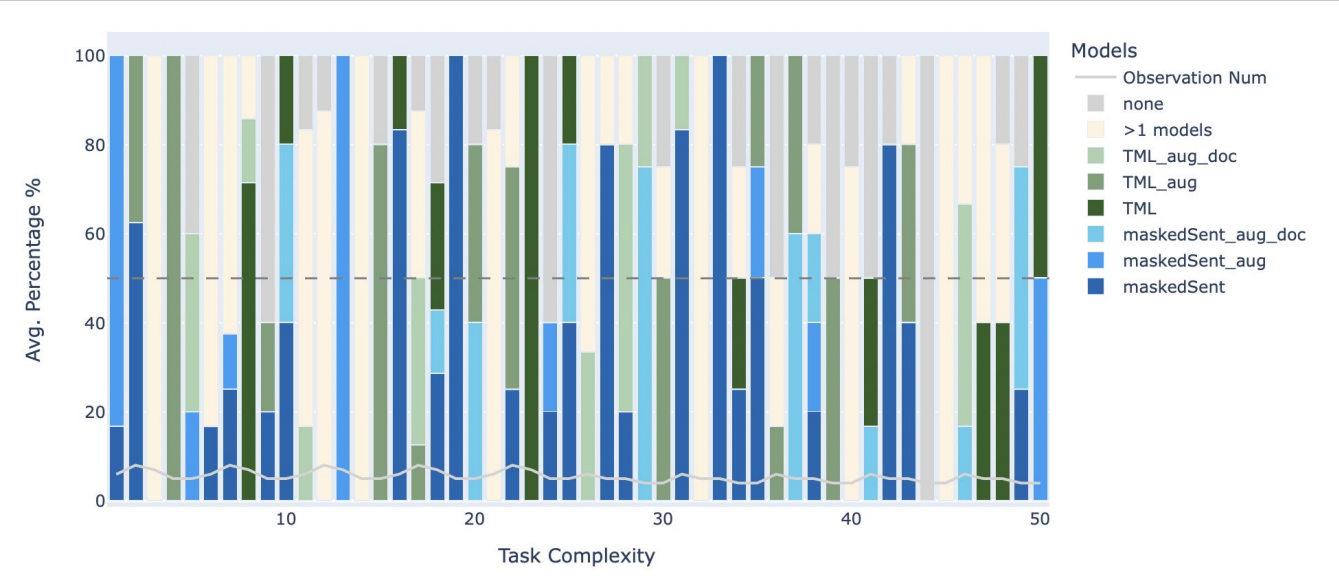


Performance comparison

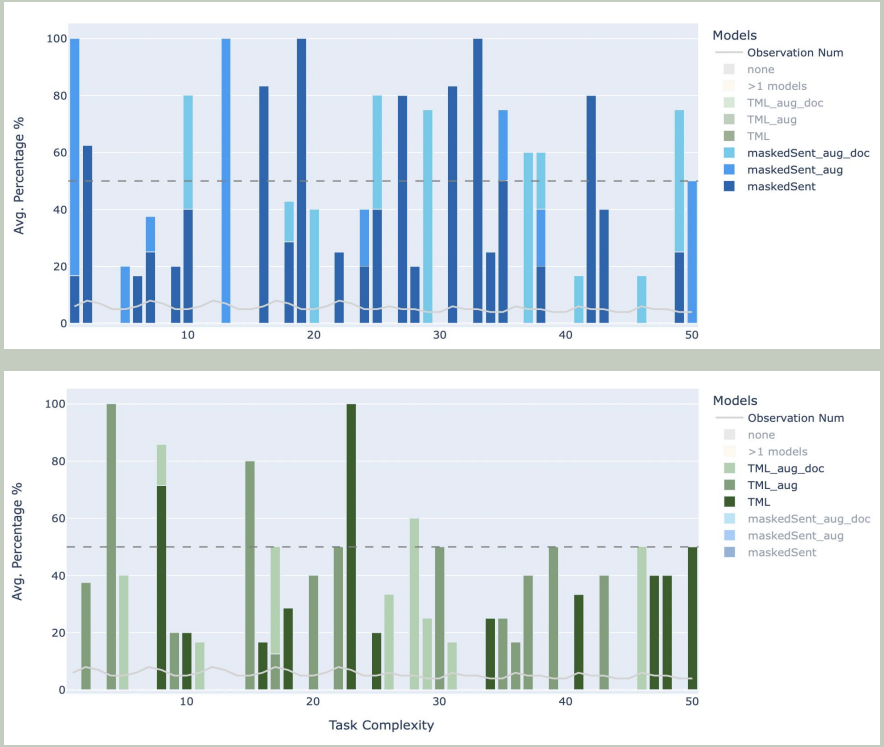
Between 6 different models

5

Distribution of Models based on Selected Labels

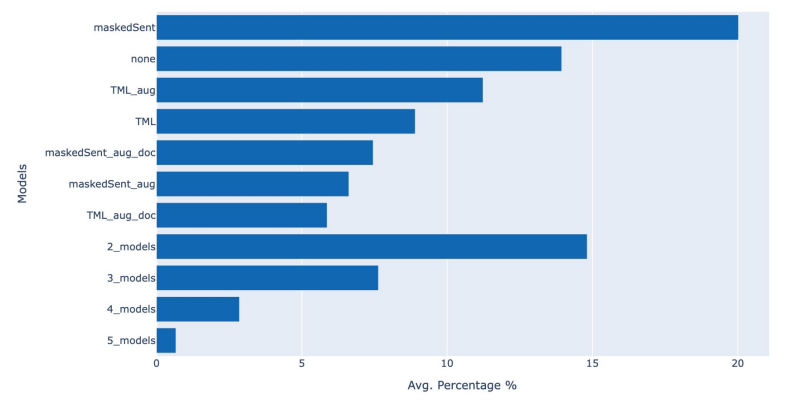


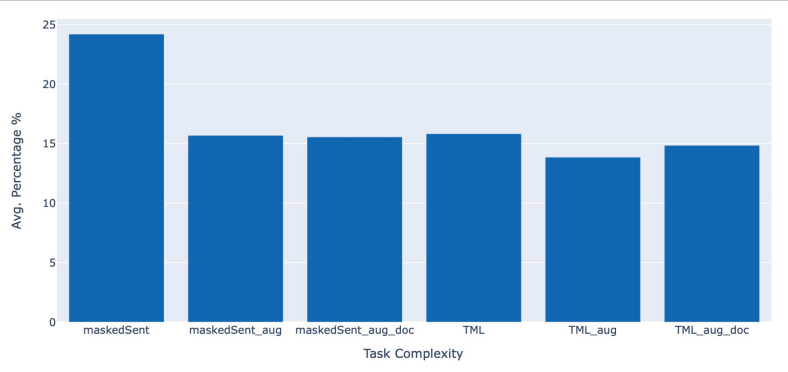
Distribution of Models based on Selected Labels



MaskedSent based models seem to outperform TML based models, but not yet leaving TML ones replaceable

Distribution of Models based on Selected Labels





25~30% of the time, >1 models produced the identical labels that got selected
(high agreement between models when the case scenario is clear)

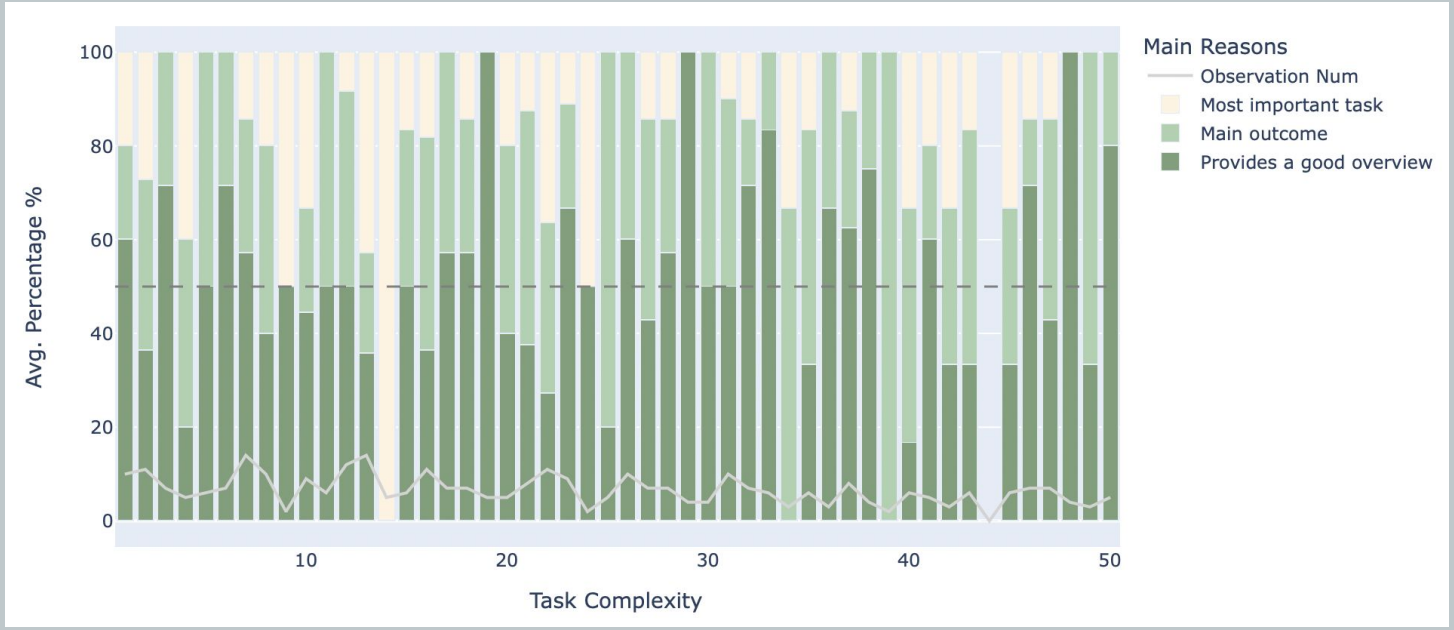
Evaluators' Main Reasons behind their label selection

Most important task, main outcome,

or good overview provided?

6

Distribution of Main Reasons across all cases

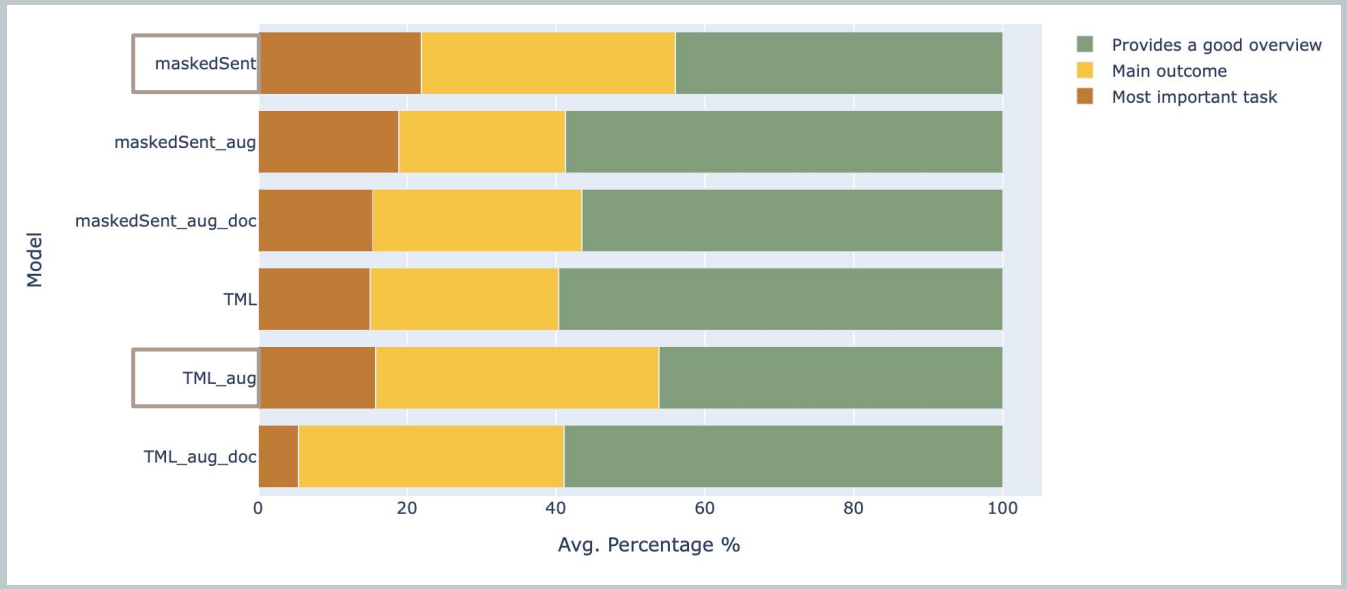


"Provides good overview", "Main outcome" are chosen in most cases, especially when task complexity is higher

Distribution of Main Reasons across all cases

Most important task	16.65%
Main outcome	33.65%
Provides a good overview	49.70%

Distribution of Main Reasons across all models



MaskedSent and TML_aug (top performing models) seem to also provide a more balanced views

Qualitative Analysis

Take a look into labeling suggestions from evaluators

7

Case 1. Extracted labels

register at online shop, select atricle, pay

Tip: In our project setting, only the labels of events (circular elements) and tasks (rectangle elements) in the process model are extracted. Due to data quality issues, not all the labels of events and tasks can be extracted successfully.

*** 125. Choose the most suitable generated label for the process presented: (Select one option)**

- shop at atricle
- register online shop
- shop online
- register account
- none of the above is relevant (if selected, please skip all questions below and go to next page)

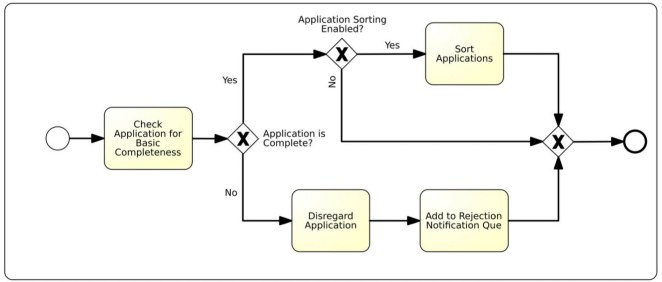
shop online

Suggestions: *'E-commerce purchase process', 'register and shop online'*

Case 2. Extracted labels

check application for basic completeness, sort applications, disregard application, add to rejection notification que

Tip: In our project setting, only the labels of events (circular elements) and tasks (rectangle elements) in the process model are extracted. Due to data quality issues, not all the labels of events and tasks can be extracted successfully.



*** 253. Choose the most suitable generated label for the process presented: (Select one option)**

- check application for basic completeness
- process application
- registration for incoming applications
- check application
- none of the above is relevant (if selected, please skip all questions below and go to next page)

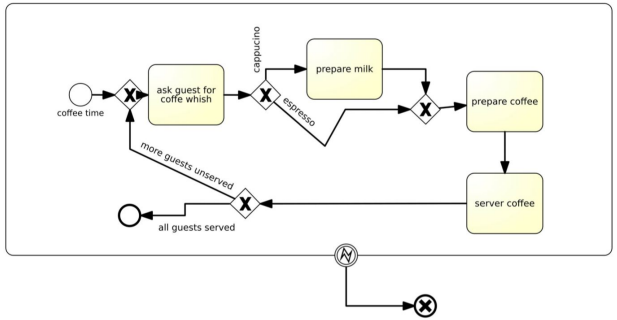
manage application received

Suggestions: *'pre-filter applications', 'Basic Application Requirements Check'*

Case 3. Extracted labels

coffee time, ask guest for coffe wish, prepare milk, prepare coffee, server coffee

Tip: In our project setting, only the labels of events (circular elements) and tasks (rectangle elements) in the process model are extracted. Due to data quality issues, not all the labels of events and tasks can be extracted successfully.



*** 198. Choose the most suitable generated label for the process presented: (Select one option)**

- serve coffee
- prepare coffee
- preparation of the menu for the event
- none of the above is relevant (if selected, please skip all questions below and go to next page)

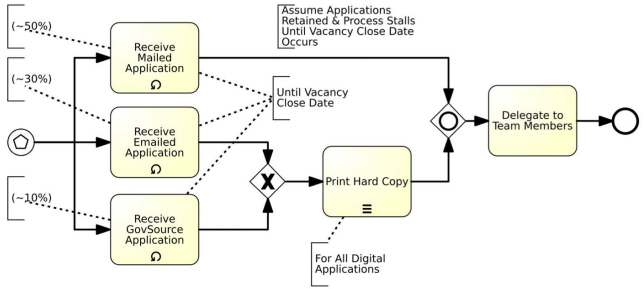
manage coffee break

Suggestions: *'prepare and serve coffee for guests during break', 'prepare coffee break'*

Case 4. Extracted labels

receive mailed application, receive emailed application, receive govsource application, print hard copy, delegate to team members

Tip: In our project setting, only the labels of events (circular elements) and tasks (rectangle elements) in the process model are extracted. Due to data quality issues, not all the labels of events and tasks can be extracted successfully.



*** 204. Choose the most suitable generated label for the process presented: (Select one option)**

- manage application
- send email reminders
- manage members
- handle members application
- none of the above is relevant (if selected, please skip all questions below and go to next page)

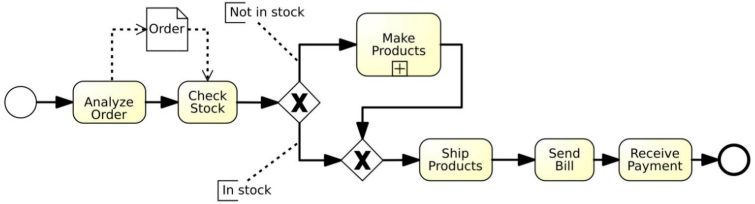
receive applications

Suggestions: *'initial management of incoming applications', 'manage applications'*

Case 6. Extracted labels

analyze order, check stock, make products, ship products, send bill, receive payment

Tip: In our project setting, only the labels of events (circular elements) and tasks (rectangle elements) in the process model are extracted. Due to data quality issues, not all the labels of events and tasks can be extracted successfully.



*** 34. Choose the most suitable generated label for the process presented: (Select one option)**

- handle order
- perform order
- verify stock and prepare shipment
- handle purchase order
- none of the above is relevant (if selected, please skip all questions below and go to next page)

fulfill order

Suggestions: *'Order-to-Cash'*, *'Fulfill order and handle payment'*

To Conclude...

8

Summary

- **Overall model performance is comparable to human**
  - General satisfaction decreases when task complexity increases
    - Data quality issue might be involved
- **Most errors can be alleviated with**
  - More labeled training data
  - Targeted learning
- **Agreement between evaluators**
  - High when selecting labels (M)
  - Lower in finer levels of evaluation (performance comparisons)
- **MaskedSent outperforms** but not (yet) able to replace the rest
- **All models provide a good overview** of the underlying process

Future work

- Account for **data quality issue**
- Address **limited data issue**
- Incorporate **contextual information** or **guided training**
- Design **continuous (re-)training pipeline** in production
  - Expert feedback loop

Main Contribution

- **Automatic generation for semantically meaningful labels for process data**
  - High-level activity labels
  - Sub-processes (in process model)
  - Process models
- **Design evaluation metrics and incorporate human evaluation**

Potential Use Cases

How can we quantify improvement?

9

Use Cases

- Label suggestions provided to **reduce manual labeling time**
- Labels provided at **different granularity level of processes**
  - Improve data management and process understanding
  - Optimize resource management
  - Increase process reusability
  - Improve the search efficiency in process knowledge base
- How do we **quantify the improvements**?
  - **Expert in the feedback loop** of model optimization allows us to **measure the revision time** throughout a period of time

Thanks!

Any questions?

You can find me at:

- yenting.wang.tw@gmail.com