Duplicate bug report detection through machine learning techniques

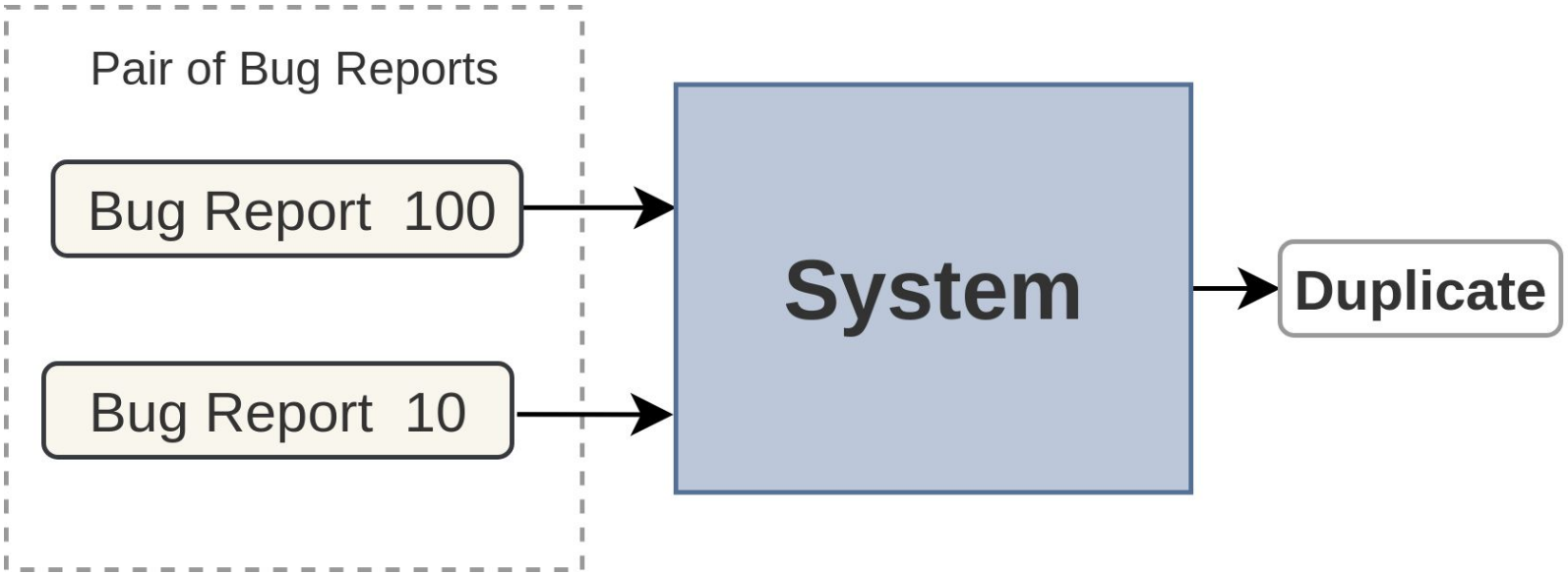
Irving Muller Rodrigues
March 10, 2018

Prof. Daniel Aloise and Prof. Michel Dagenais

Polytechnique Montréal
Laboratoire **DORSAL**

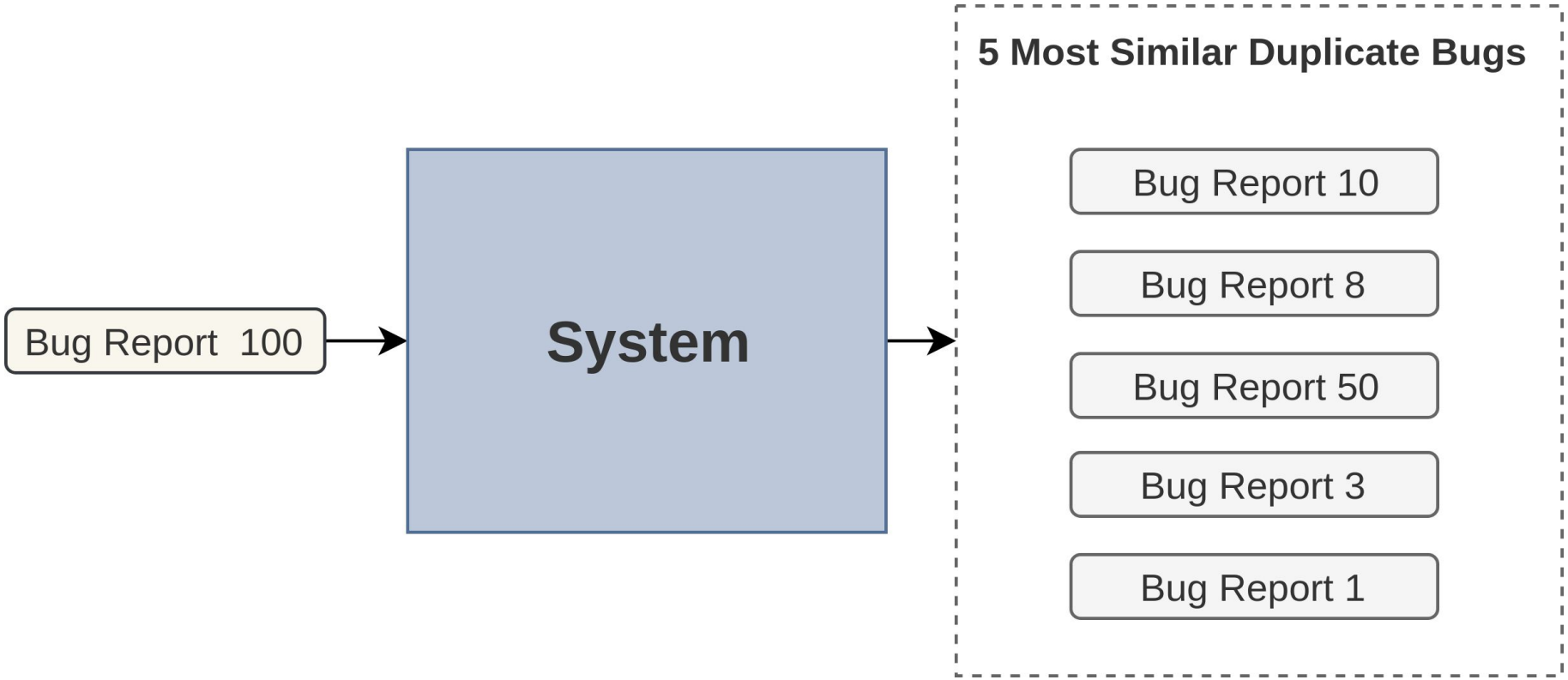
Approaches

- Decision-making approach
  - Training and test dataset have pairs of bug reports



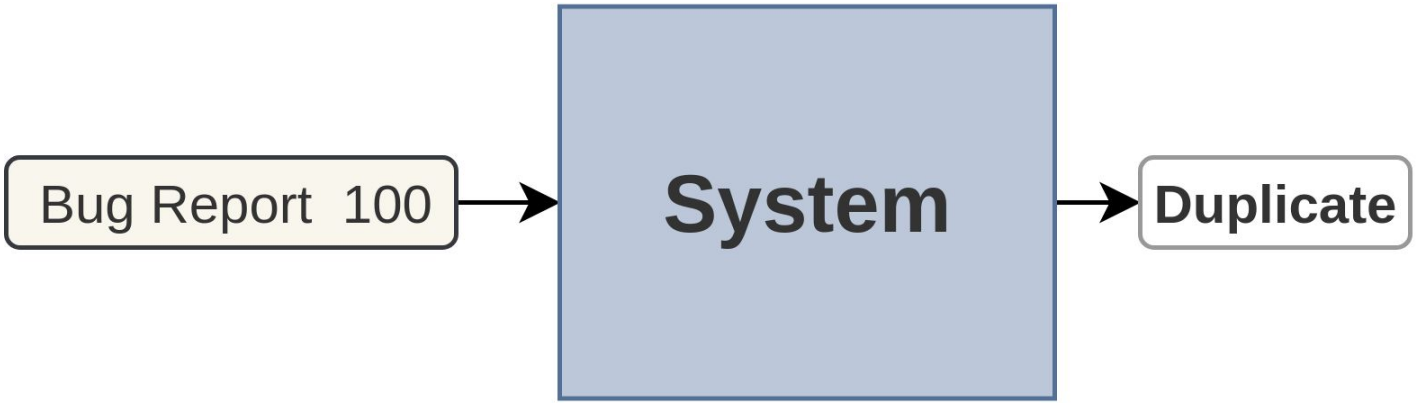
Approaches

- Ranking approach



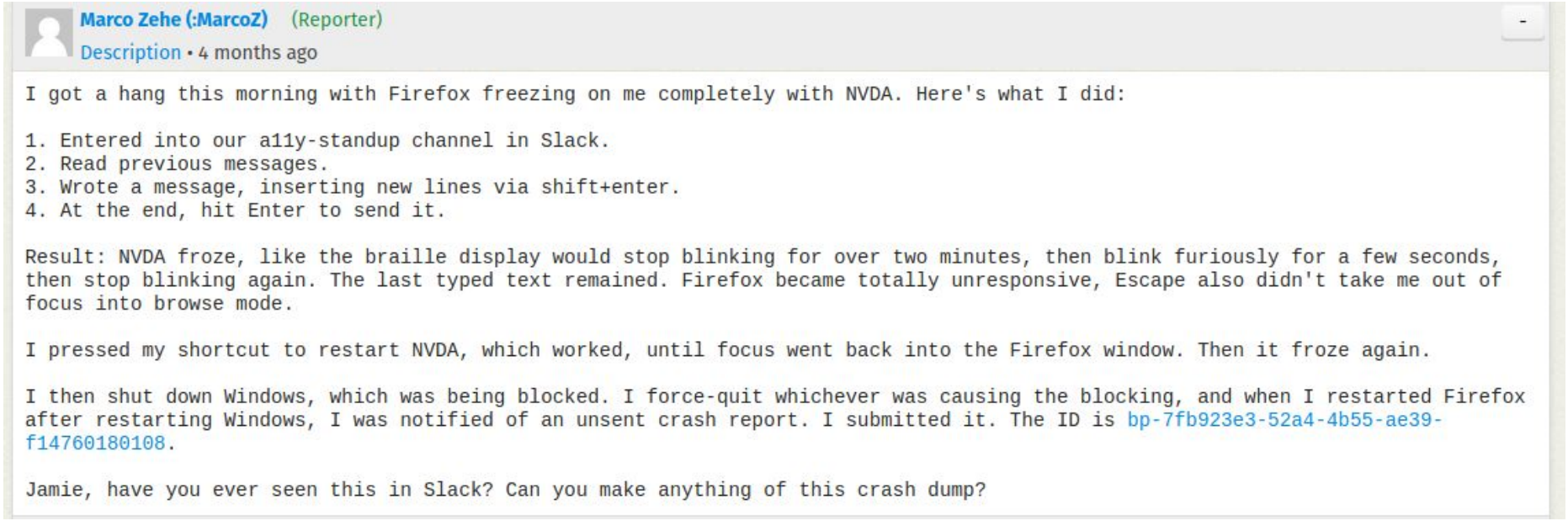
Approaches

- Binary classification approach
  - Classifier input: usually general information extracted from the database and the new bug reports



Heterogeneous Information

- Textual information



Marco Zehe (:MarcoZ) (Reporter)

Description • 4 months ago

I got a hang this morning with Firefox freezing on me completely with NVDA. Here's what I did:

1. Entered into our a11y-standup channel in Slack.
2. Read previous messages.
3. Wrote a message, inserting new lines via shift+enter.
4. At the end, hit Enter to send it.

Result: NVDA froze, like the braille display would stop blinking for over two minutes, then blink furiously for a few seconds, then stop blinking again. The last typed text remained. Firefox became totally unresponsive, Escape also didn't take me out of focus into browse mode.

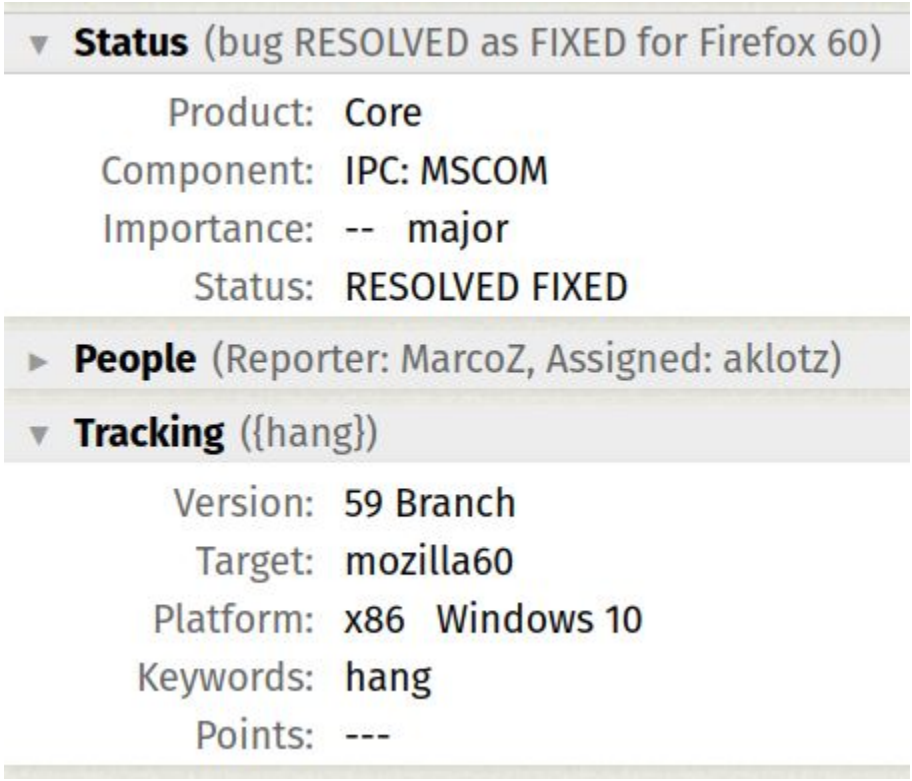
I pressed my shortcut to restart NVDA, which worked, until focus went back into the Firefox window. Then it froze again.

I then shut down Windows, which was being blocked. I force-quit whichever was causing the blocking, and when I restarted Firefox after restarting Windows, I was notified of an unsent crash report. I submitted it. The ID is bp-7fb923e3-52a4-4b55-ae39-f14760180108.

Jamie, have you ever seen this in Slack? Can you make anything of this crash dump?

Heterogeneous Information

- Categorical information



Heterogeneous Information

- Stack traces

```
eclipse.buildId=I200405211200
An internal error occurred during: "Debug".
java.lang.ArrayIndexOutOfBoundsException: 1
org.eclipse.jface.util.ListenerList.add(ListenerList.java:104)
org.eclipse.jface.viewers.LabelProvider.addListener(LabelProvider.java:43)
org.eclipse.debug.internal.ui.LazyModelPresentation.getPresentation
(LazyModelPresentation.java:176)
org.eclipse.debug.internal.ui.LazyModelPresentation.getText
(LazyModelPresentation.java:101)
org.eclipse.debug.internal.ui.DelegatingModelPresentation.getText
(DelegatingModelPresentation.java:158)
org.eclipse.debug.internal.ui.views.DebugViewLabelDecorator$LabelJob.run
(DebugViewLabelDecorator.java:326)
org.eclipse.core.internal.jobs.Worker.run(Worker.java:66)
```

Heterogeneous Information

- Logs

```
[ RUN      ] PersonalDataManagerTest.GetProfileSuggestions_InvalidData
../../components/autofill/core/browser/personal_data_manager_unittest.cc:1969:
Failure
Expected equality of these values:
  base::ASCIIToUTF16("1234567890")
    Which is: 1234567890
  suggestions[0].value
    Which is: 9876543210
Stack trace:
../../components/autofill/core/browser/personal_data_manager_unittest.cc:1970:
Failure
Expected equality of these values:
  base::ASCIIToUTF16("9876543210")
    Which is: 9876543210
  suggestions[1].value
    Which is: 1234567890
======
```

Heterogeneous Information

- Code

```
onkeydown=function(){
    window.open('//example.com/','_blank','a');
}

onkeypress=function(){
    window.open('about:blank','_blank').close();
}
```

- Much more types of information!

Heterogeneous Information

- Classical Approach
  - Feature Engineering
  - Handcrafted features for each type of information
  - Time consuming and expensive
  - How to combine them?
  - Curse of dimensionality

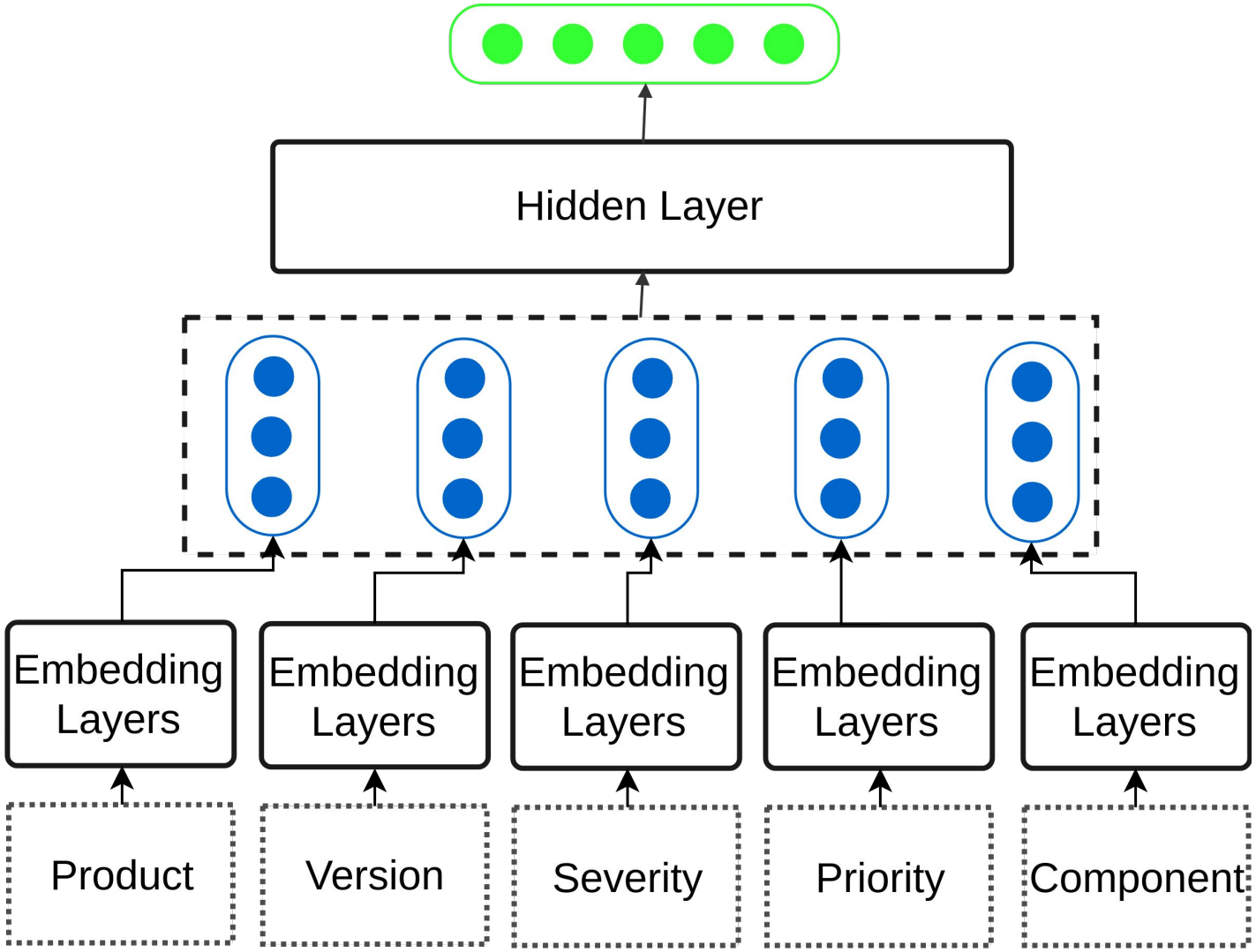
Heterogeneous Information

- Deep Learning
  - Feature learning automatically
  - Encode information in a vector
  - End-to-end systems
  - Tools to encode information:
    - Recurrent Neural Network
    - Autoencoder
    - Embedding
    - Adversarial Nets
    - etc...

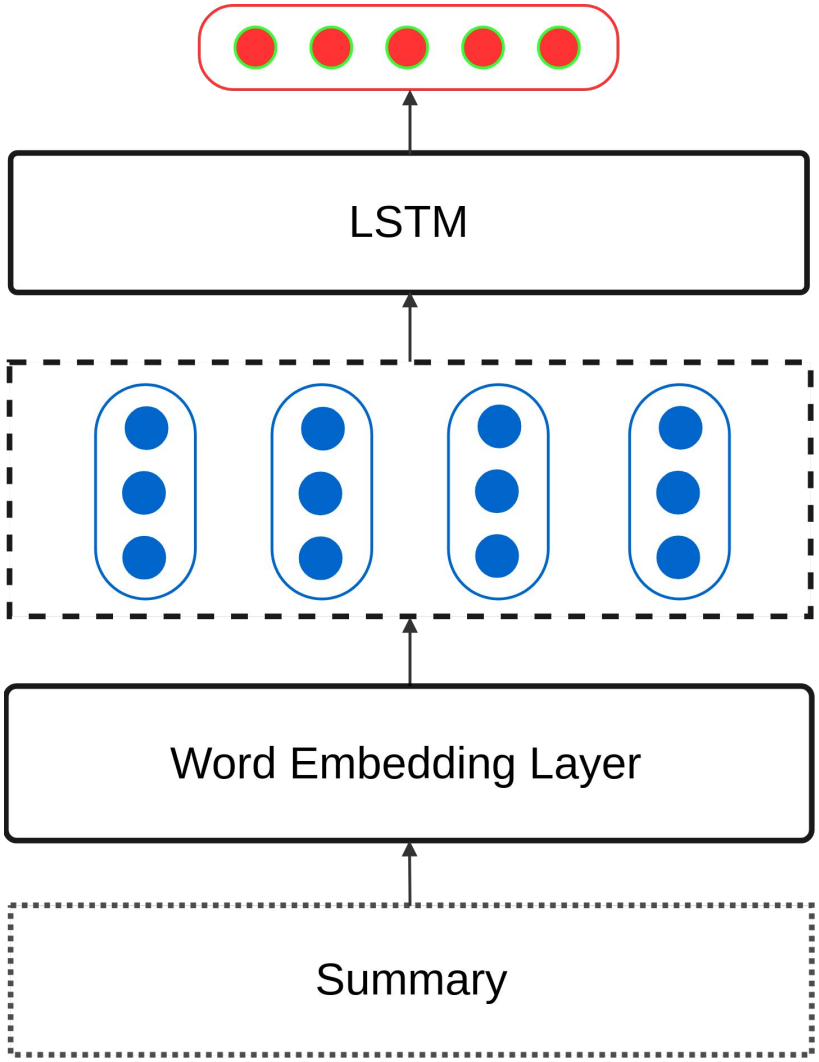
Experiment

- Decision-making approach
- Model
  - Based on Deshmukh et al. 2017
  - Input: a pair of bug reports
  - Categorical information
    - Severity, product, version, priority, component
    - Encoded by Meta Encoder
  - Summary information
    - Encoded by Summary Encoder

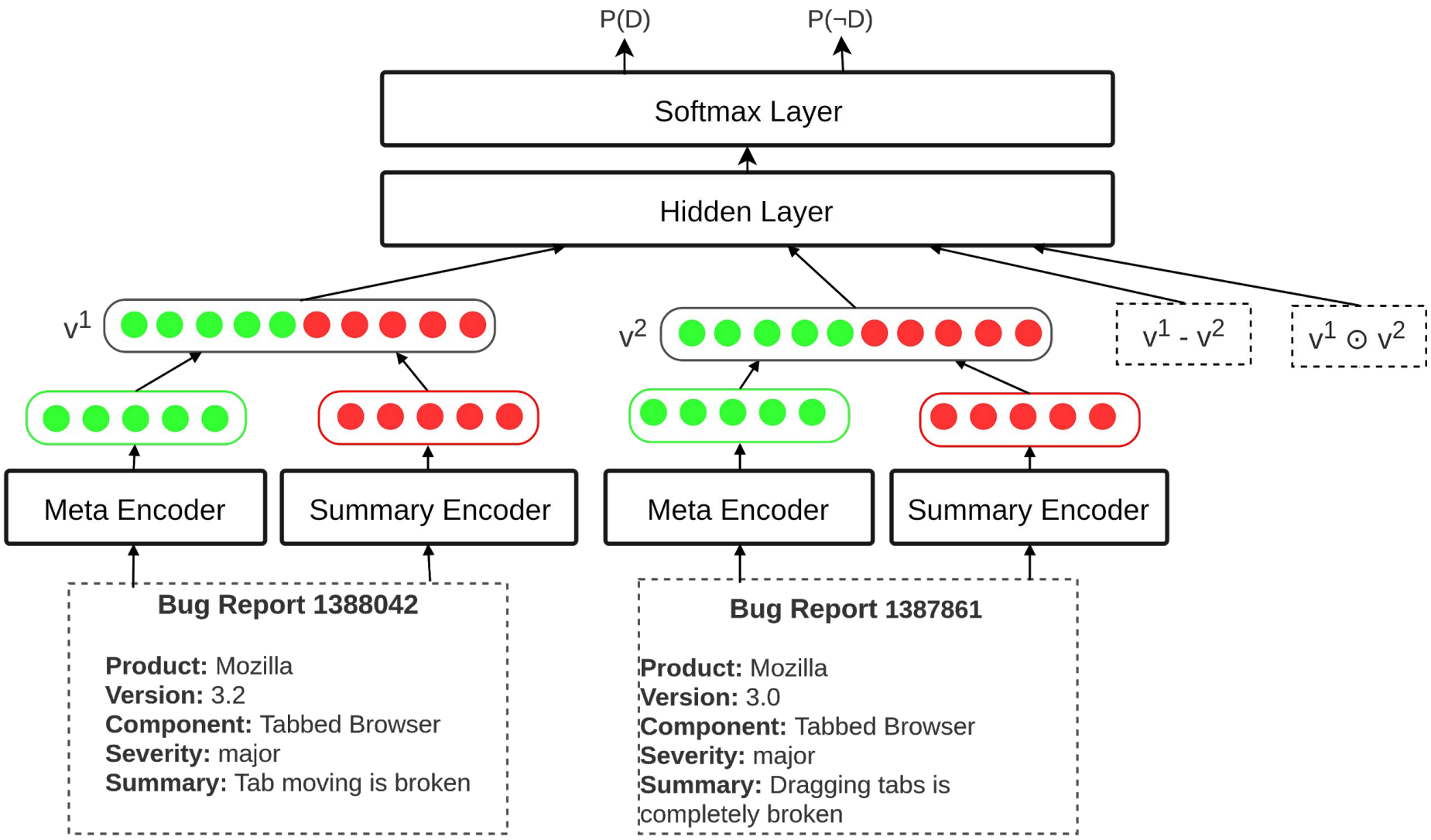
Meta Encoder



SummaryEncoder



Model


P(D)
P(¬D)
Softmax Layer
Hidden Layer
v1
v2
v1 - v2
v1 ⊙ v2
Meta Encoder
Summary Encoder
Meta Encoder
Summary Encoder
Bug Report 1388042
Product: Mozilla
Version: 3.2
Component: Tabbed Browser
Severity: major
Summary: Tab moving is broken
Bug Report 1387861
Product: Mozilla
Version: 3.0
Component: Tabbed Browser
Severity: major
Summary: Dragging tabs is completely broken

Experiment

- Differences of our model and Deshmukh et al. 2017
  - Use only categorical and summary information
  - Fine-tune our word embedding during the training
  - Compute difference and multiplication wise of vectors
  - Filters: module and function
    - Regex
  - Their best model calculates the cosine distance
    - Tune threshold

Experiment

- Dataset
  - Data from Lazar et al. 2014
  - Pairs of bug reports
  - 3 different domains: Eclipse, Open Office and Net Beans
  - Pair of Non duplicate bugs were randomly generated
  - 50% pair of duplicate bugs : 50% pair of non duplicate bugs

Distribution of pairs

Domínio	Train	Test
OpenOffice	177,793	44,449
Eclipse	176,289	44,073

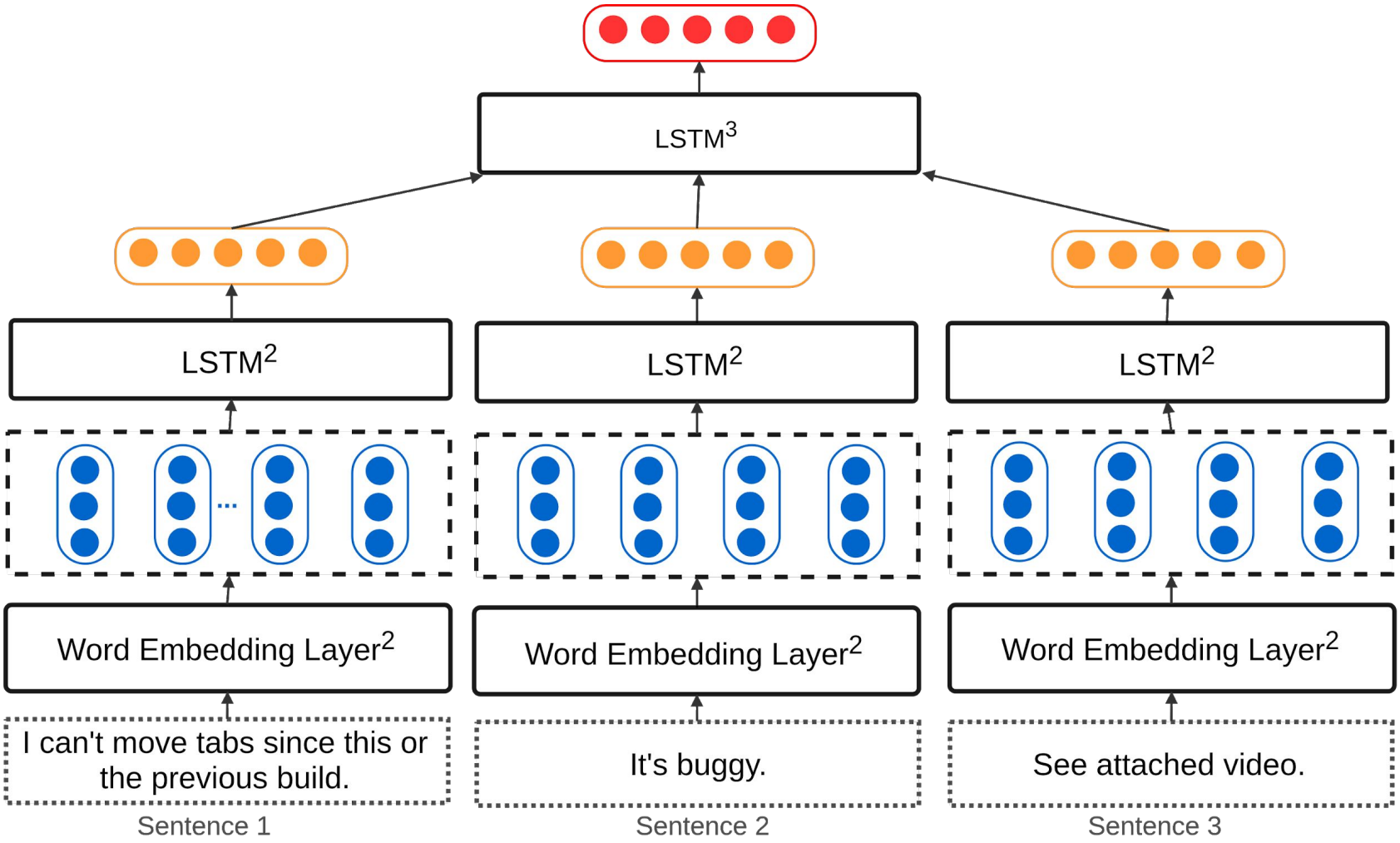
Preliminar Results

Domínio	Our Model	Deshmukh et al. 2017
OpenOffice	95.67	94.55
Eclipse	95.99	90.60

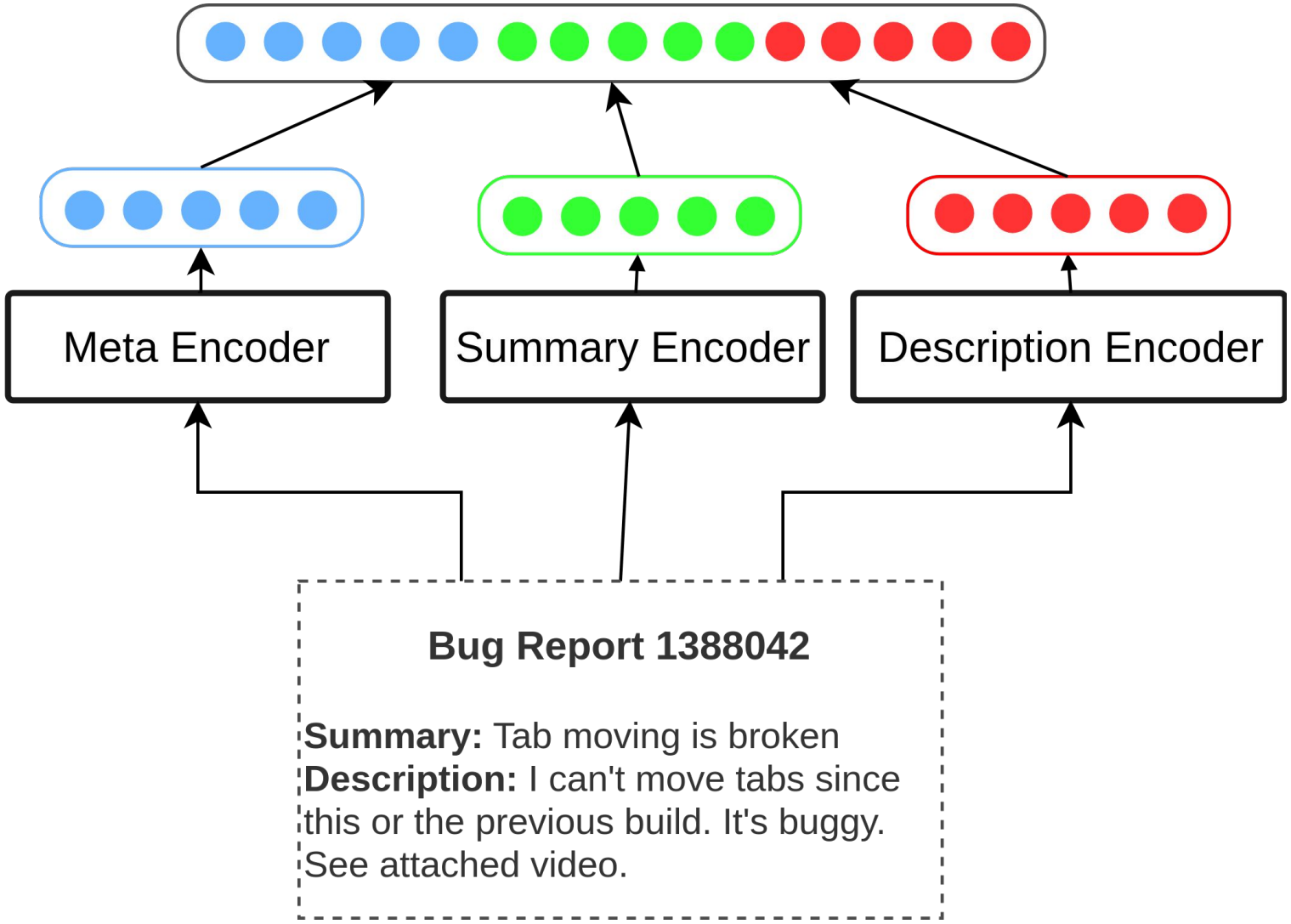
Future Directions

- Description Information
  - Critical: how to cleaning it?
    - Stack trace
    - Code
    - Typos and informal language
  - Description Encoder
    - Multi-layer lstm
    - Convolution neural network

Description Encoder



Description Encoder



Future Directions

- Evaluate our method using Ranking approach
  - Recall Rate
  - More useful than decision-making approach
- Evaluate on unbalanced dataset
  - More realistic scenario
  - 1:4 proportion
- Stack traces, Trace information, Code, Logs

Thank you for your attention!

Questions?

Irving Muller Rodrigues
irving.muller-rodrigues@polymtl.ca

References

- Deshmukh, J., M, A. K., Podder, S., Sengupta, S., & Dubash, N. (2017). Towards Accurate Duplicate Bug Retrieval Using Deep Learning Techniques. 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), 115–124. http://doi.org/10.1109/ICSME.2017.69
- Lazar, A., Ritchey, S., & Sharif, B. (2014). Generating duplicate bug datasets. Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014, 392–395. http://doi.org/10.1145/2597073.2597128