Week 4

Detecting and Defining Specialized Language in Text

Valerie Lopez



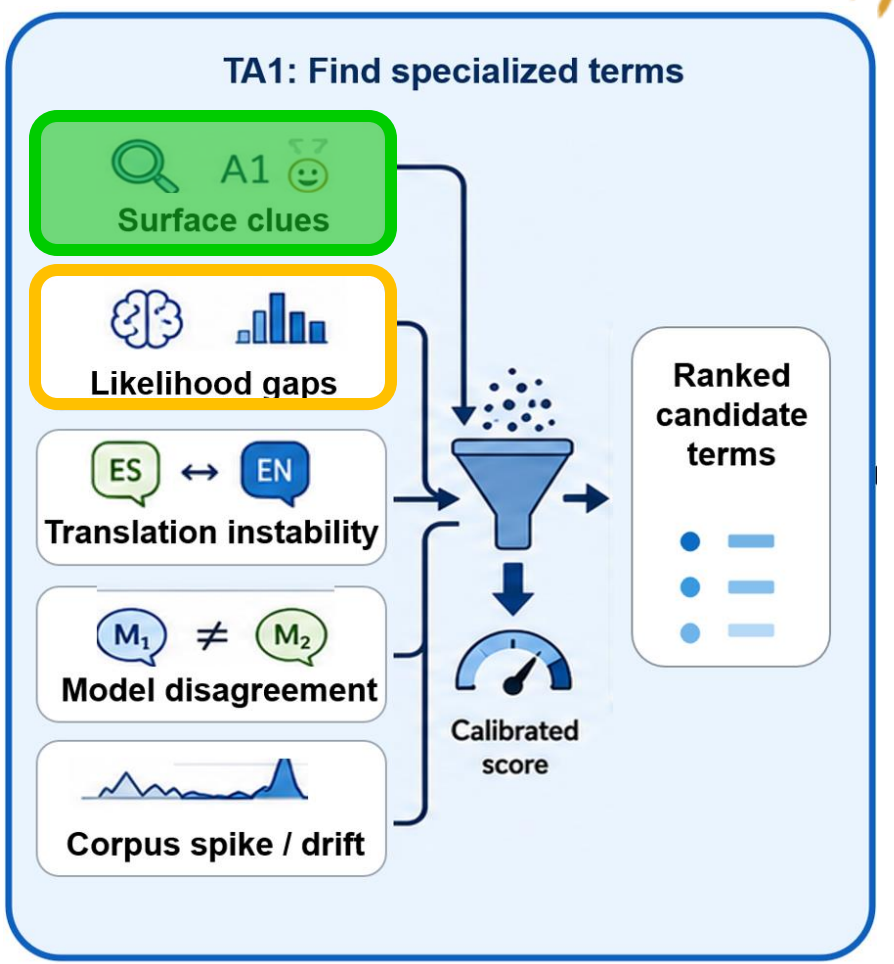
Quick Recap

Project Goals

- Identify specialized language **candidates**
- **Score** candidates using **4 detection methods**

Method 2: Language Surprise

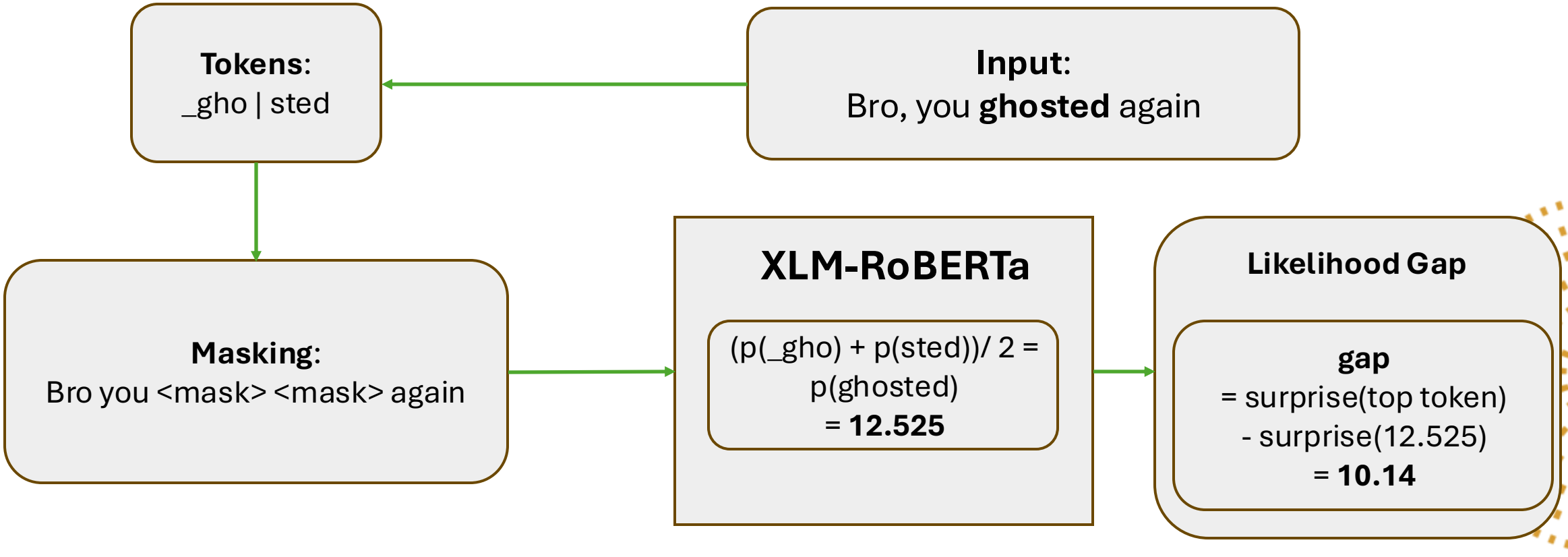
- **XLM-RoBERTa**
- Fill Mask to predict candidate **surprise**
- Calculating **likelihood gaps**



Overview: Language Surprise

- **XLM-RoBERTa** - Encoder finetuned on 100+ languages
- **Mask candidates** (words & phrases)
  - Phrases built on permutations again
- Calculate **surprise**
- Calculate **gap**

Lang Surprise: Pipeline



Interpreting Likelihood Gap

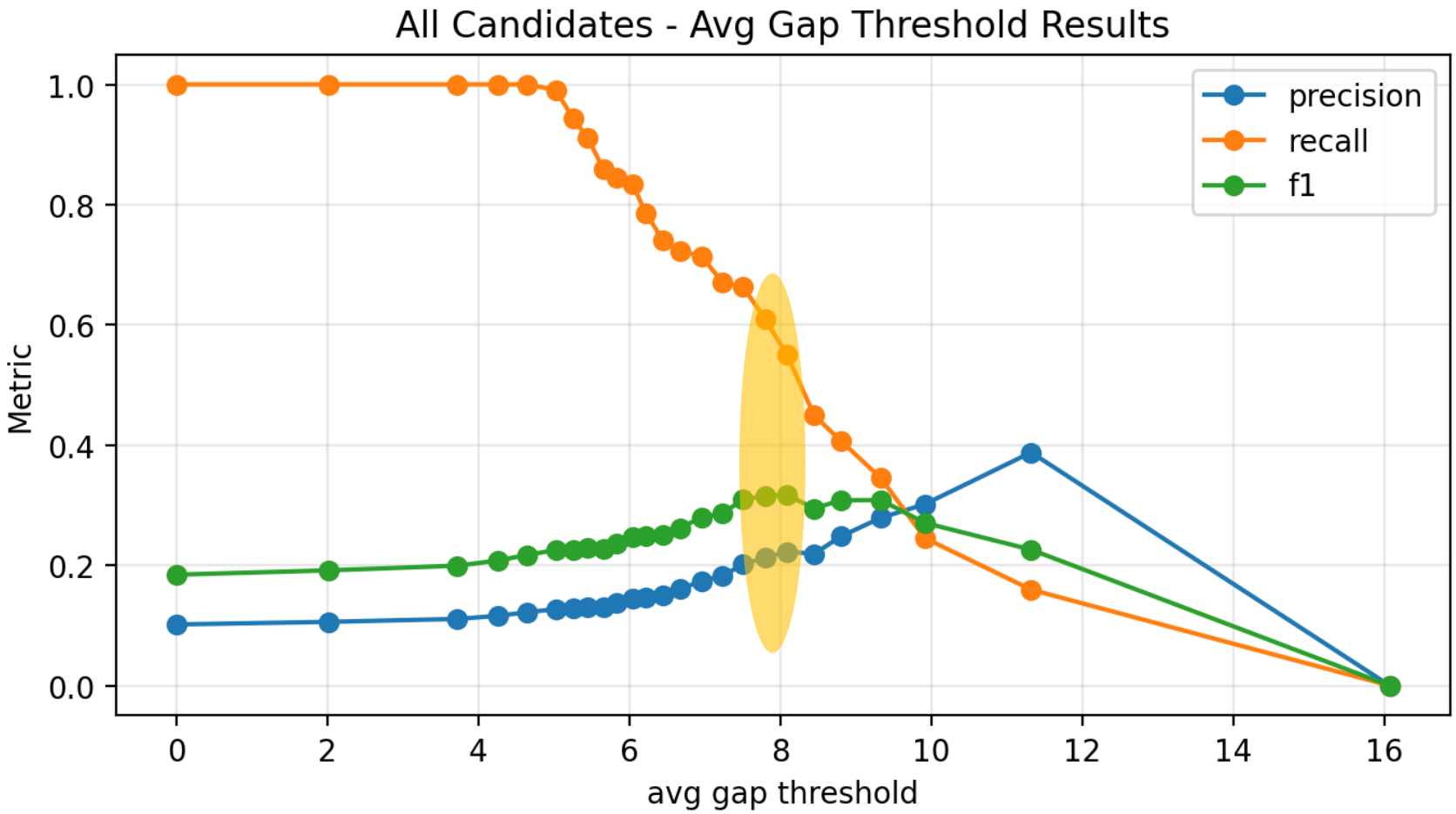
Large Gap

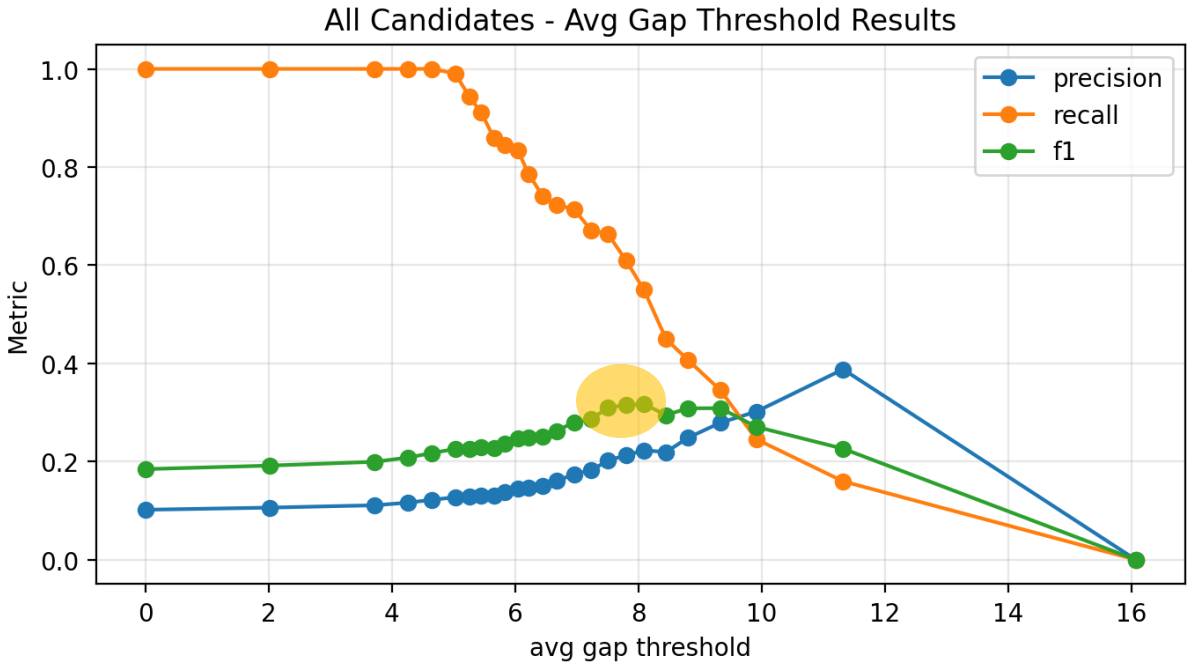
- Model highly confident
- Actual token violated expectations
- Candidate could be unusual

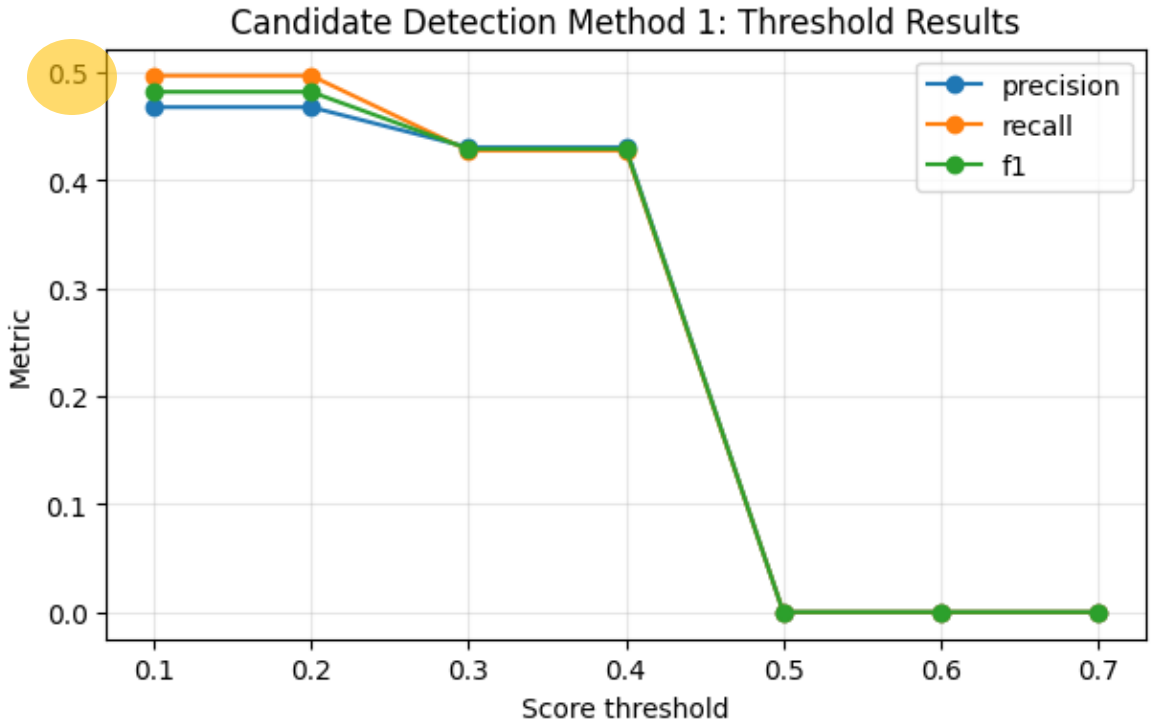
Smaller Gap

- Model uncertain
- Actual token slightly less likely than alternative
- Candidate could be ordinary

Results: Likelihood Gap







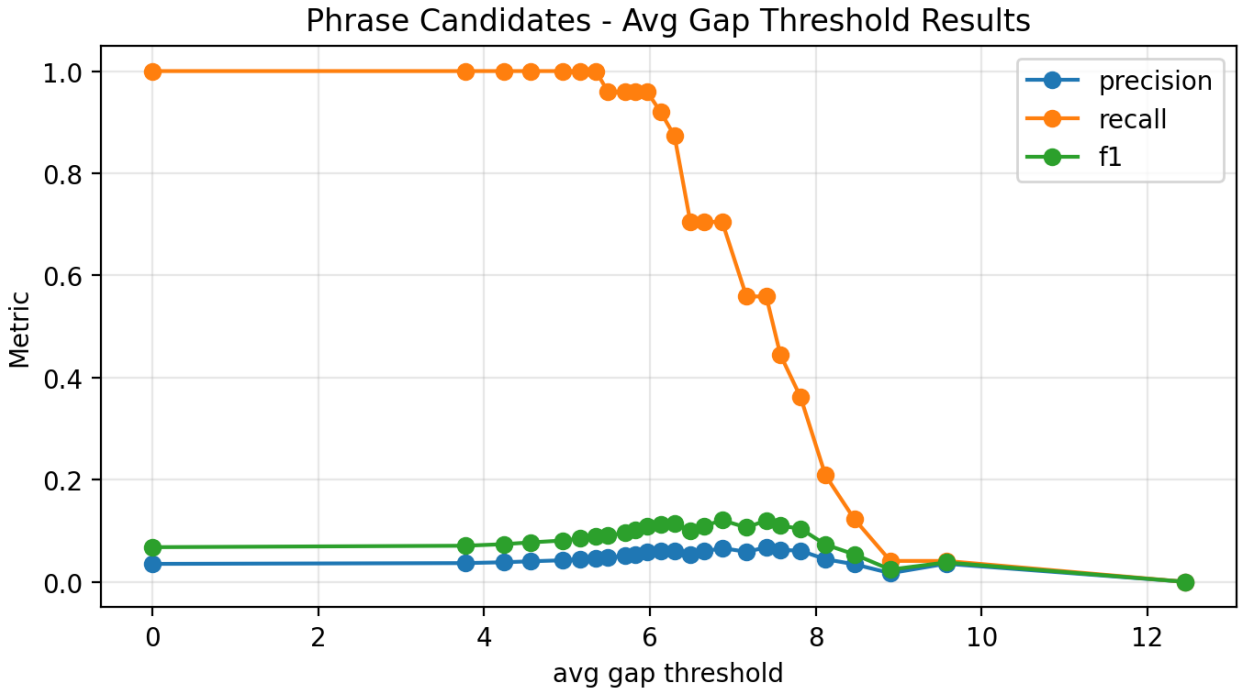
Method Performances

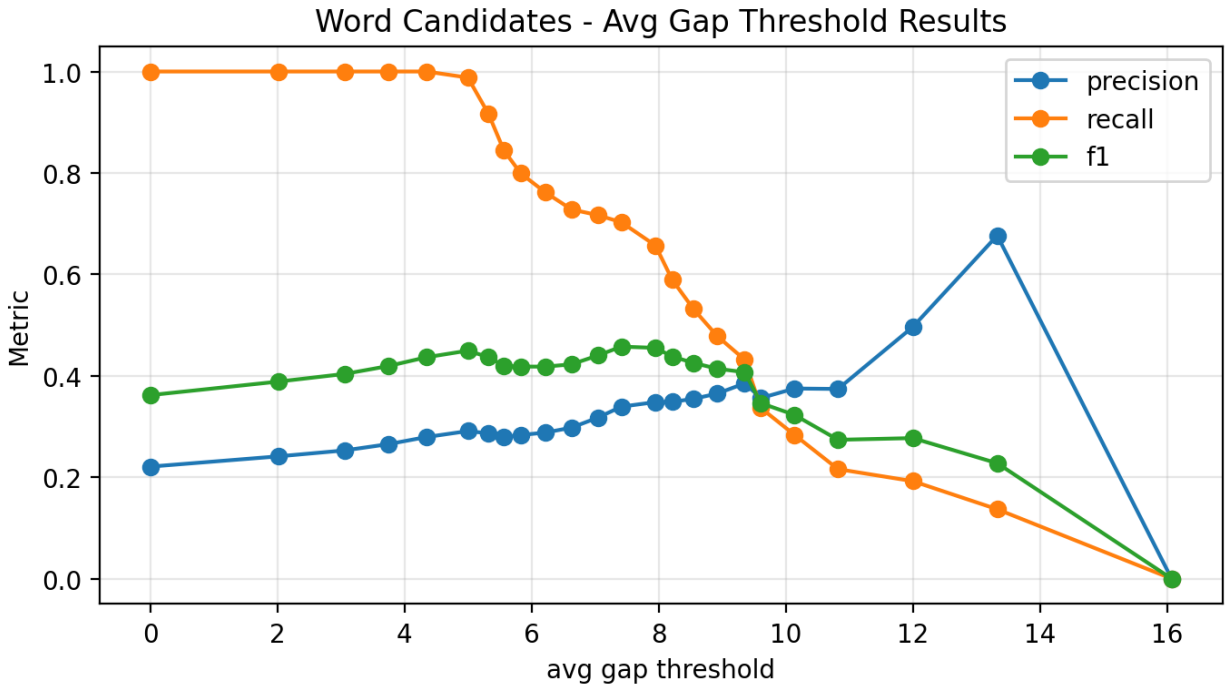
Likelihood Gap vs Surface Clues

*Corpus score removed from Method 1

Why did Method 2 perform worse?

- **Nature of the dataset**
  - Slang favors surface clues
- **Language surprise fails with odd sentences**
  - High FP rate for regular terms
- **Gap inflation** from non-meaningful top tokens
  - Masking approach is faulty



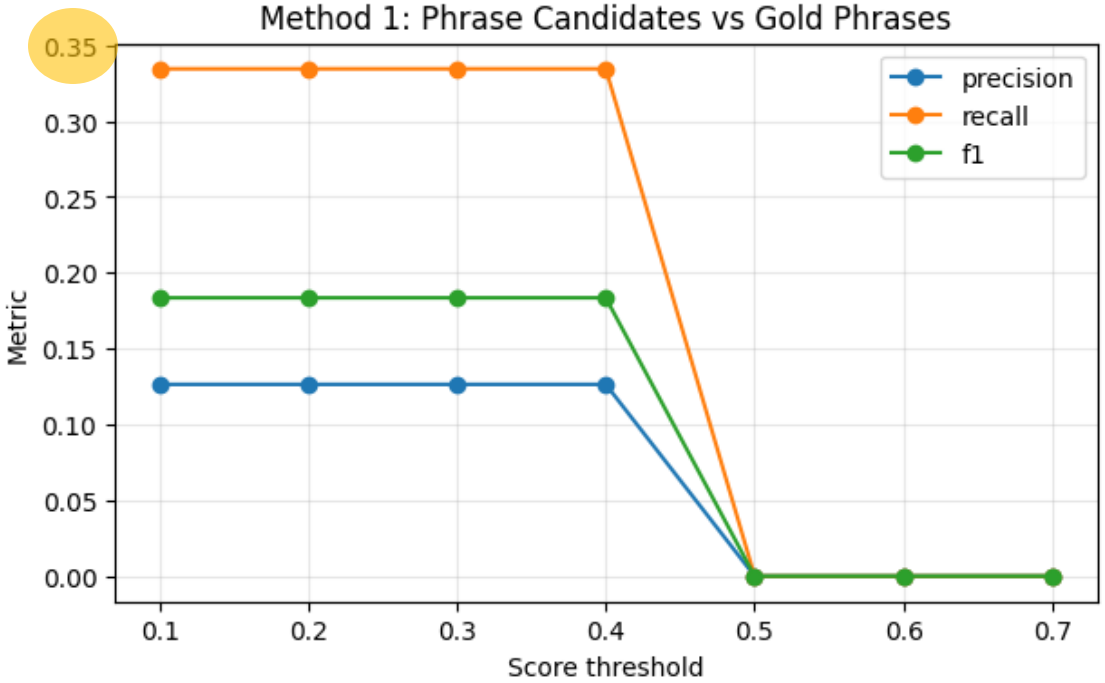


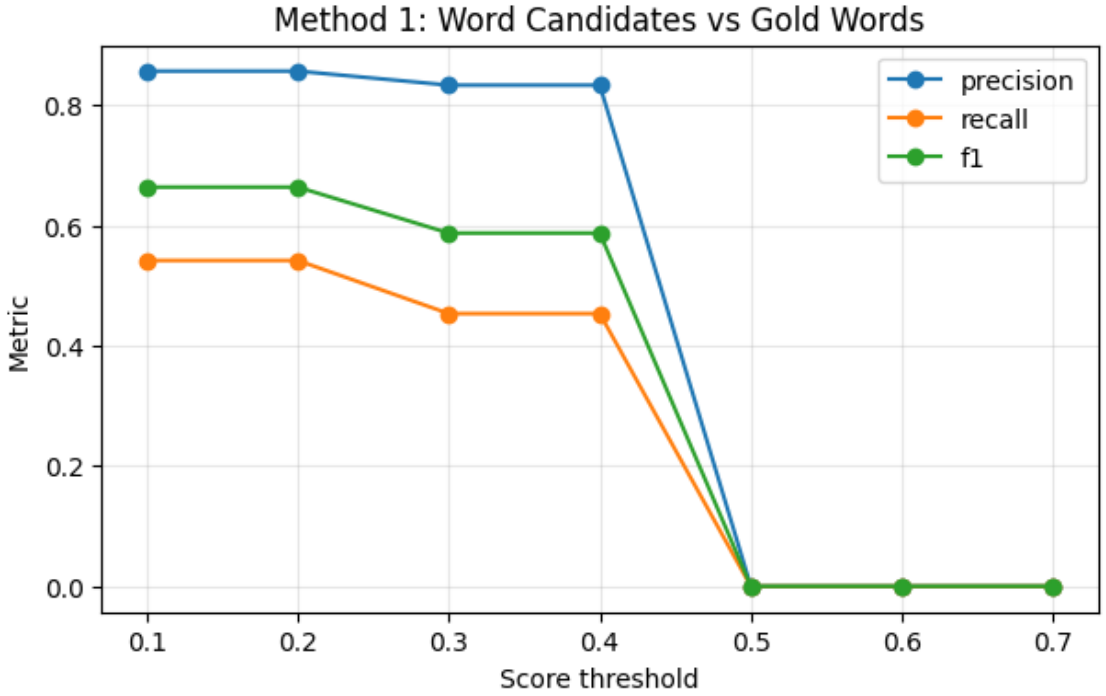
Experiment: Candidate Type vs Performance

Context:

Surface clues performed poorly due to phrases

- **Increase in FP** rate leads to **low precision** for phrases





Experiment: Candidate Type vs Performance

Does the theory hold for Method 1?

- **Words perform better** regardless of method

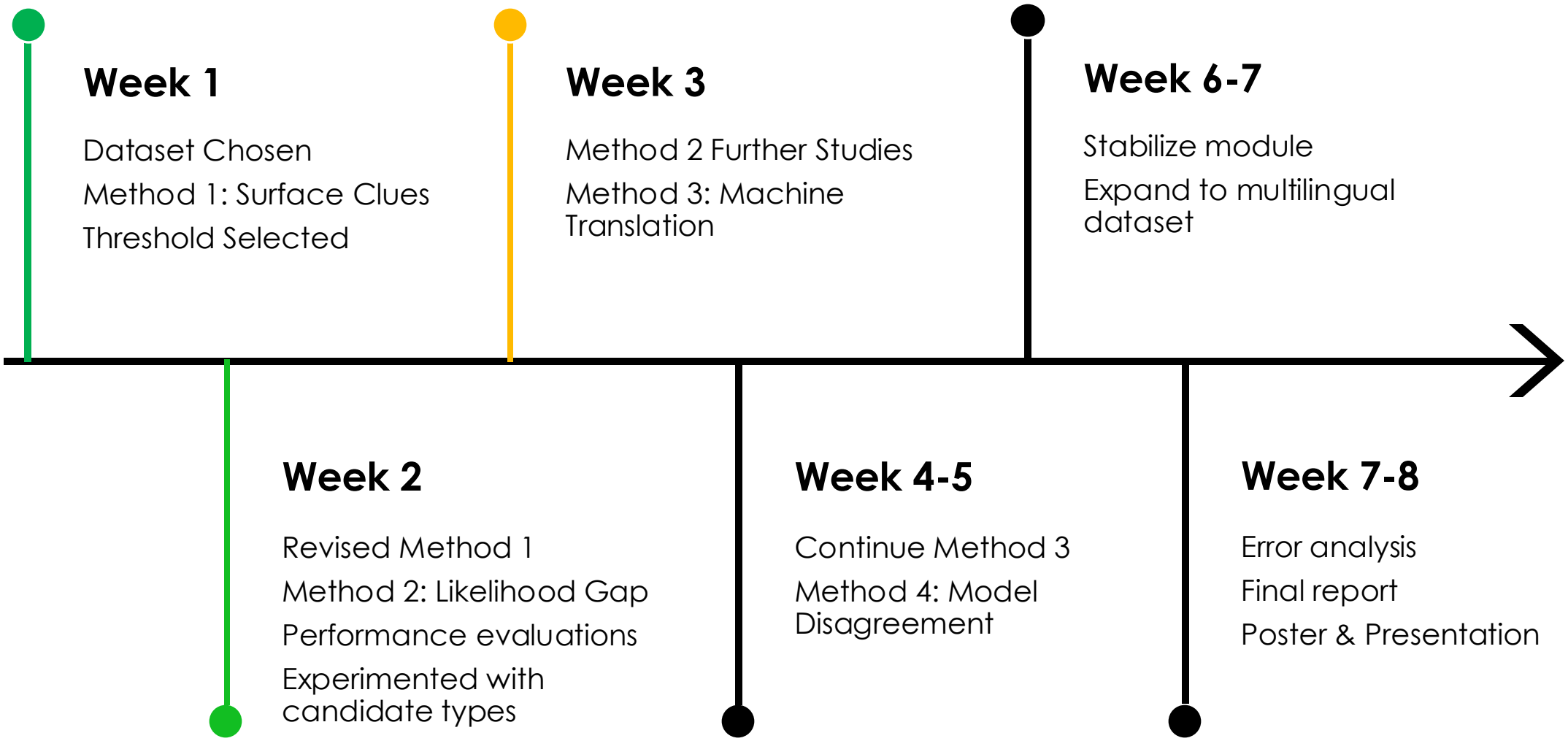
Results: Likelihood Gap

candidate set	score category	threshold	precision	recall	f1
all candidates	avg gap	8.089285	0.222725	0.54971	0.317009
phrase candidates	avg gap	6.871894	0.065974	0.704719	0.120653
word candidates	avg gap	7.411401	0.339121	0.702708	0.457471

Method 2: Further Studies

- Can we optimize our phrase generation method?
- Does masking tokens differently improve performance?
- Will a pipeline from method 1 improve method 2?

Updated Timeline



Thank You

Any Questions?

By: Valerie Lopez

