Week 5

Detecting & Defining Specialized Language in Text

Dr. Santu Karmaker, Dr. Song Wang, Dr, Mubarak Shah, Adam Bawatneh

Valerie Lopez

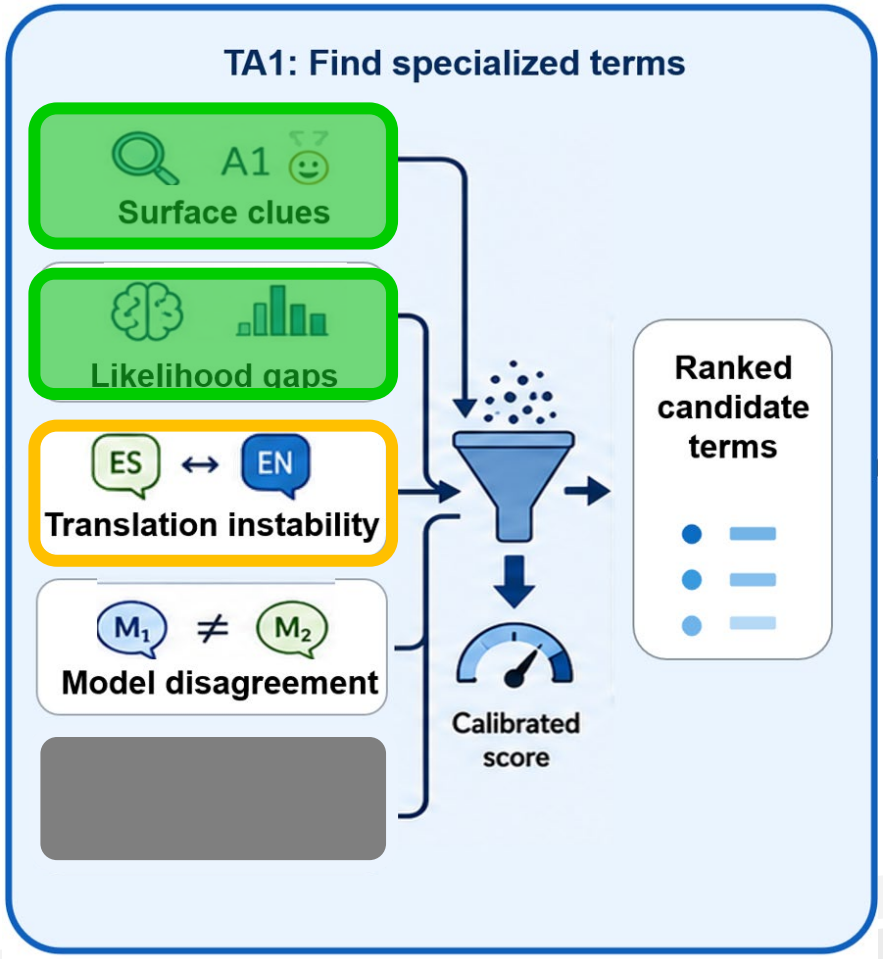


Project Recap

- Identify **special language** from text
- Apply **4** detection **methods**
- **Dataset:** acader/genz-alpha-slangs

This Week:

- Method 3: **Machine Translation**
- **Meta Llama 3.1 8B Instruct**



Method 3: Machine Translation

Input (English)	**Bruh**, **IRL** you **fold** faster than a Walmart chair
Translation (Spanish)	Hombre, en la vida real, doblas más rápido que una silla de Walmart.
Translation (English)	**Geez**, **in real life**, you can **spin** around faster than a Walmart chair.

Detected Terms: *Bruh, IRL, fold*

Fuzz Ratio

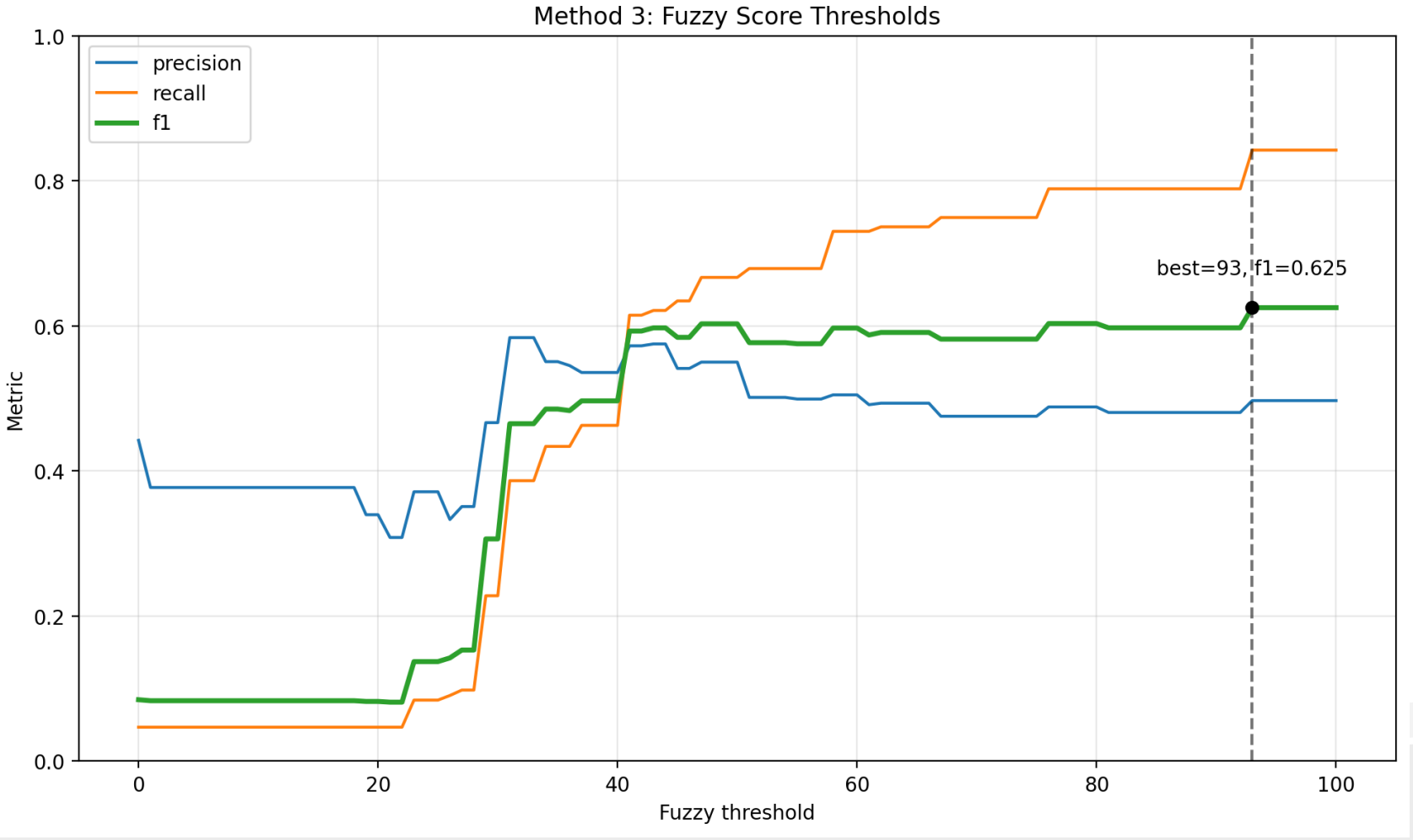
- Similar strings have higher scores
- Detects when a term **changes** or **disappears**
- **Threshold = 93**
- If **score < threshold**, terms not similar enough. **Classify as specialized!**

Input (English)
Bruh, IRL **you fold** faster than a Walmart chair

Translation (English)
Geez, in real life, **you** can spin **around** faster than a Walmart chair.

candidate	best_back_	fuzz_ratio
you	you	100
fold	around	40

Fuzzy Threshold Analysis



Named Entity

If a term is **not translated** at all and is **not a named entity**, it is specialized

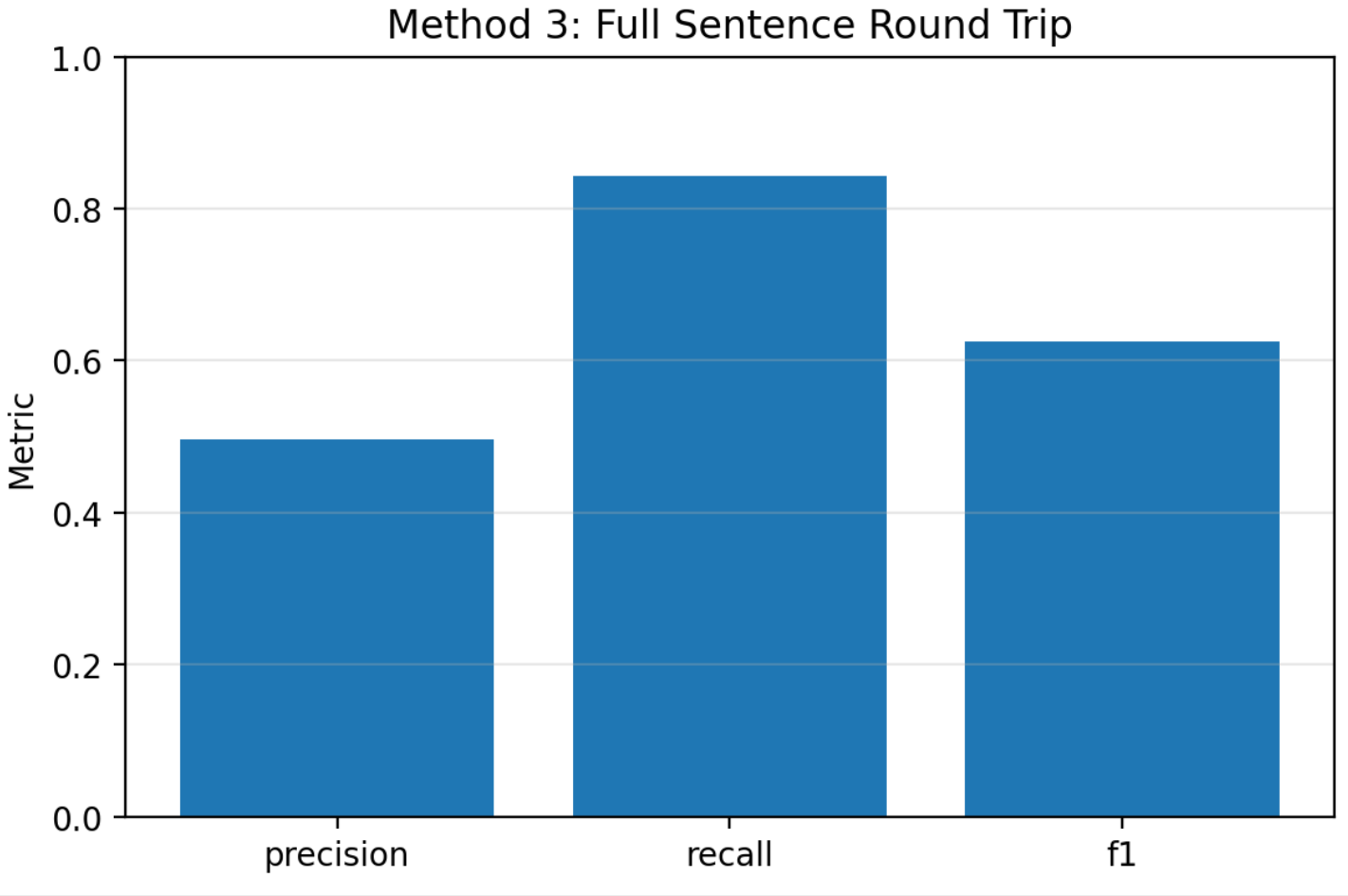
Use **NER** to parse these cases

Input (English)	Bro, **skibidi** outta here, **sigma** style
Translation (Spanish)	Hermano, **skibidi** fuera de aquí, estilo **sigma**
Translation (English)	Brother, **Skibidi** out of here, **Sigma** style

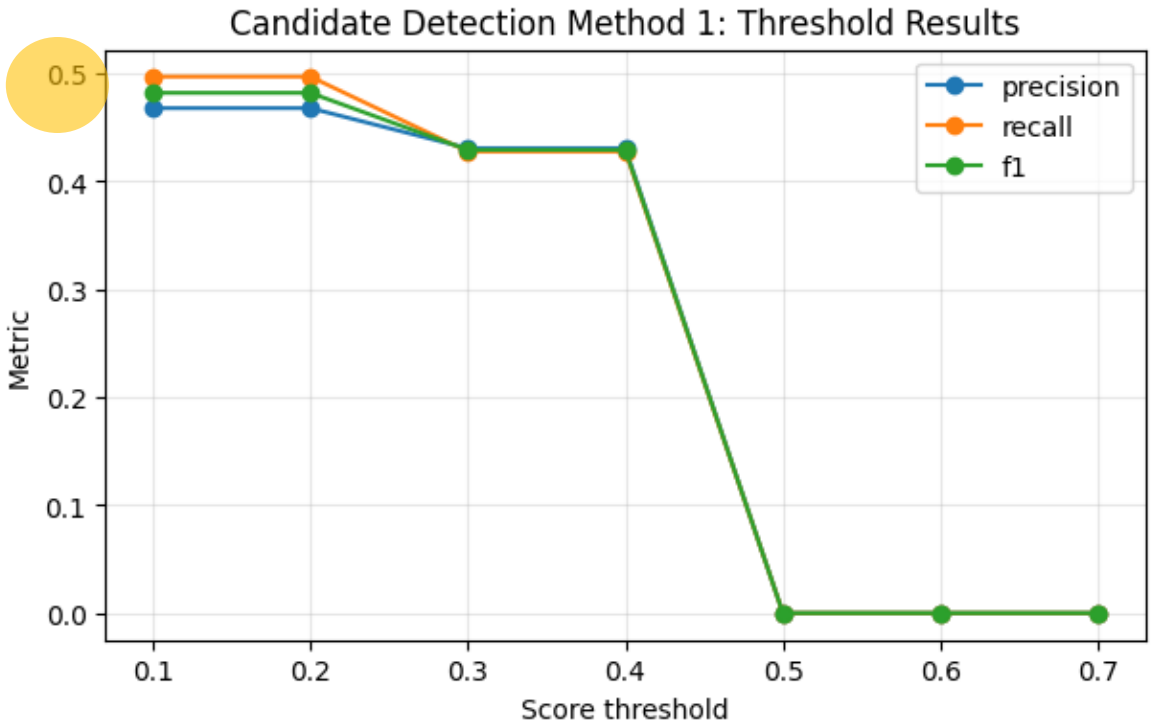
Method 3: Overview

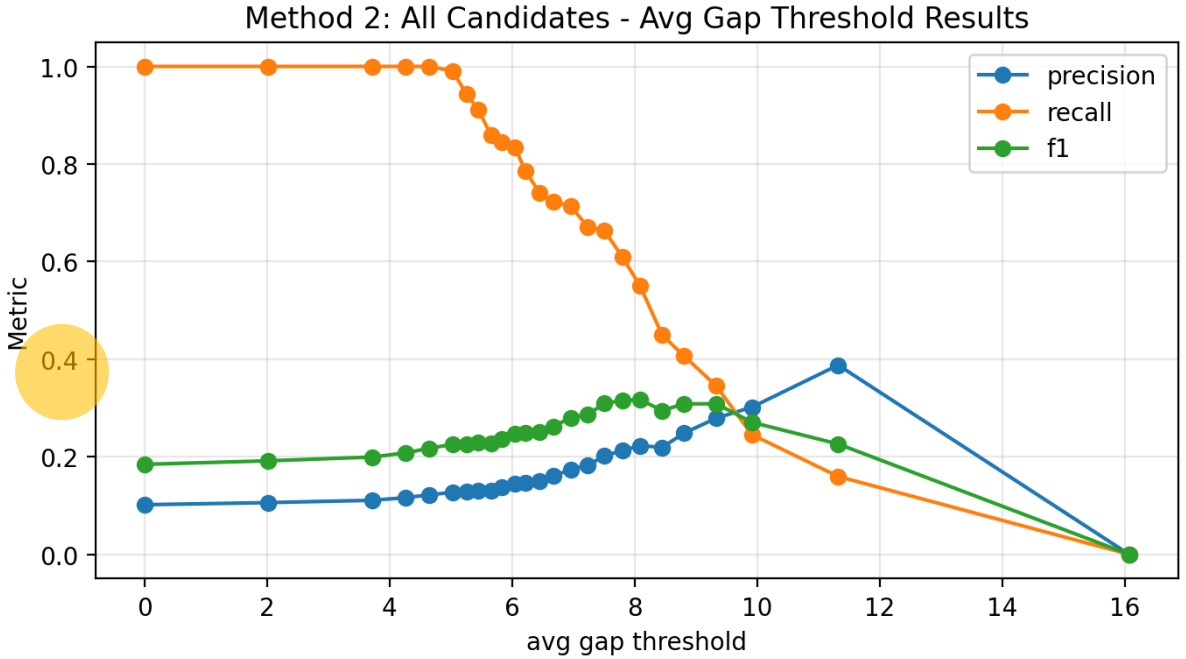
1. Given sentence input
2. Translate full sentence, **English to Spanish**
3. Translate sentence, **Spanish to English**
4. **Tokenize** original sentence
5. **Tokenize** final sentence
6. For **each token in the original sentence**
    1. If it did not translate, use **NER**
    2. Use **fuzz ratio**
        1. Compare to every **token in final sentence**, looking for highest similarity
        2. If **score low**, token likely disappeared or changed. **Conclude specialized!**

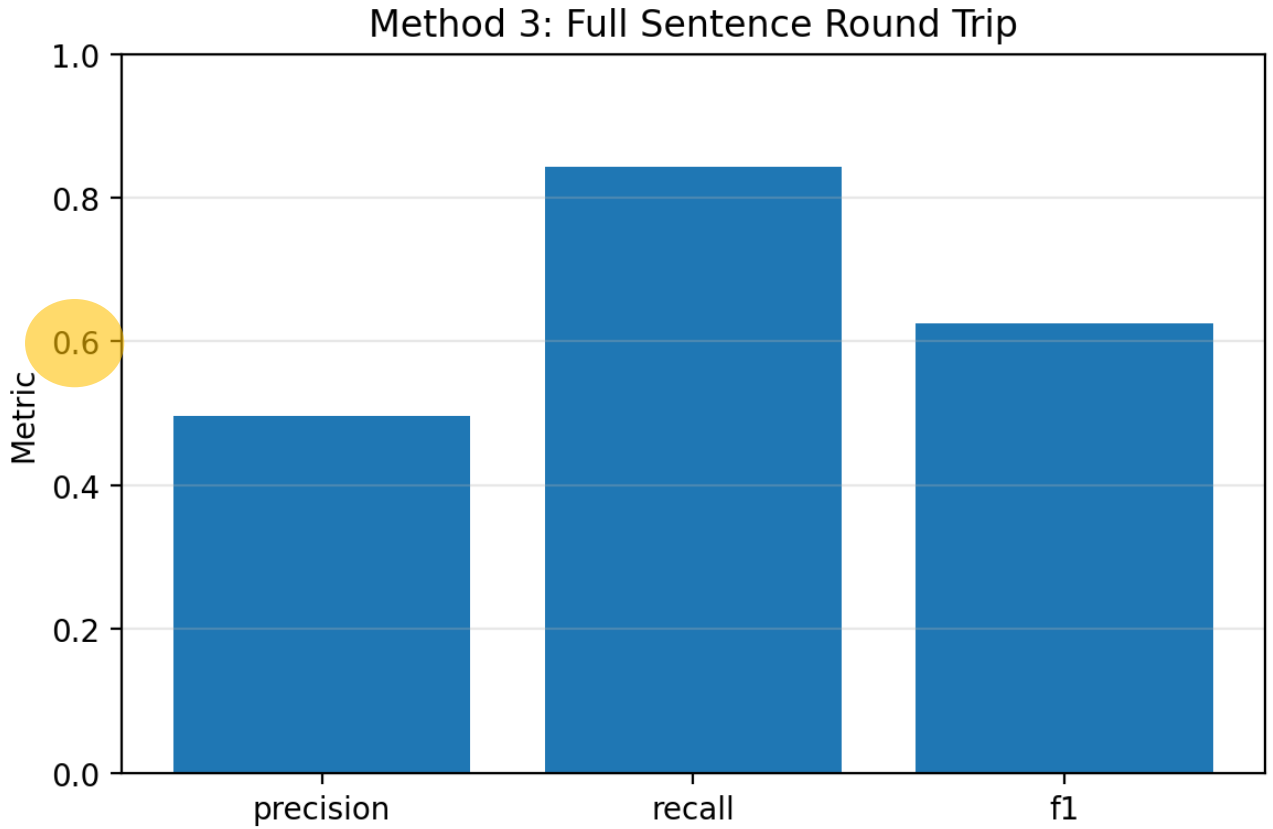
Machine Translation Performance



Method Performances

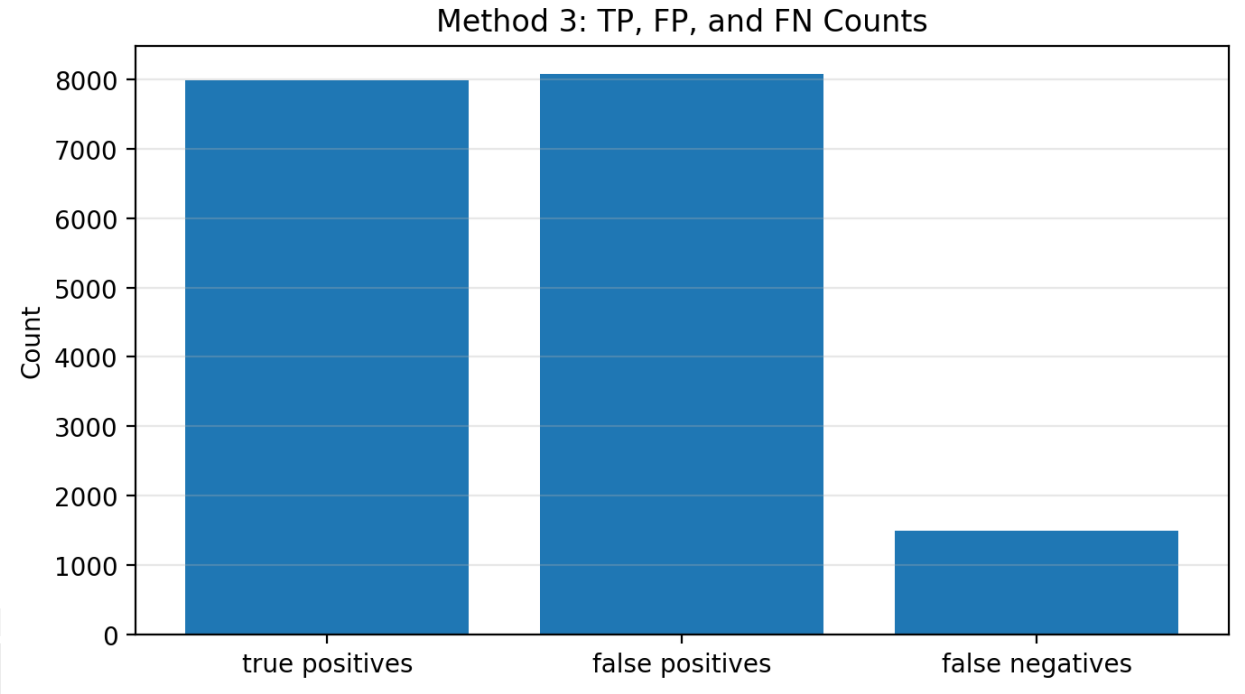




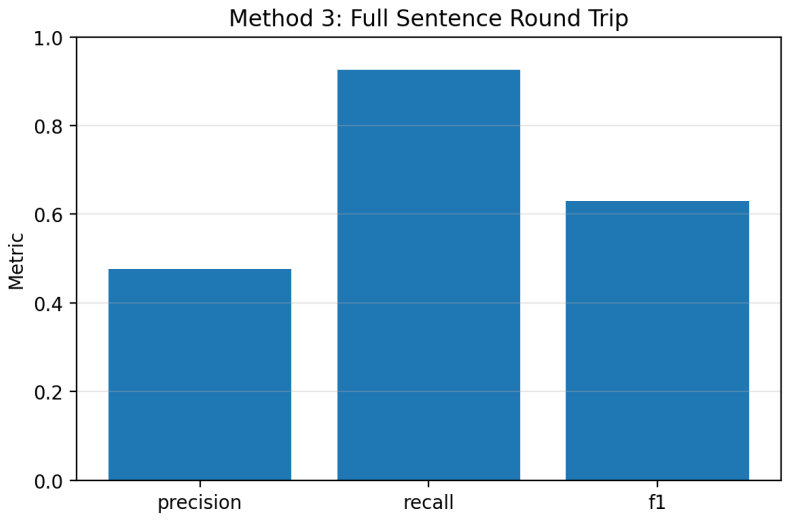


Explaining Performance

- Slang varies by language, favoring Method 3
- Dataset incorrectly labeled increases FP

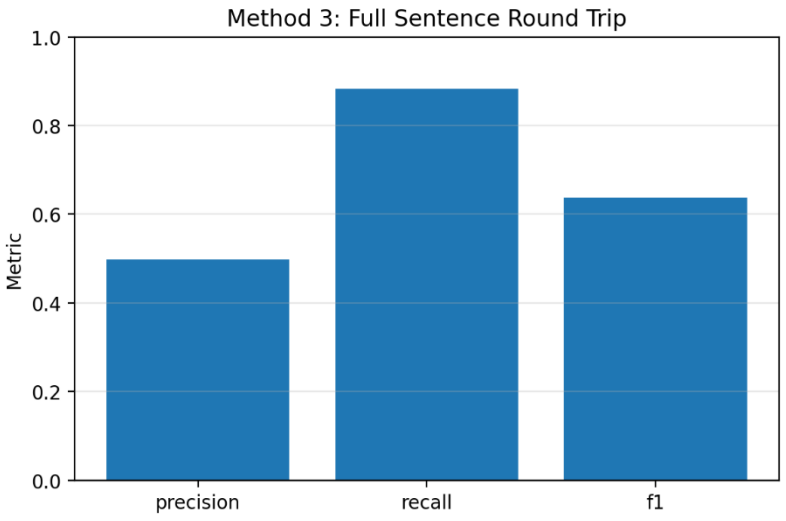


Expanding Languages



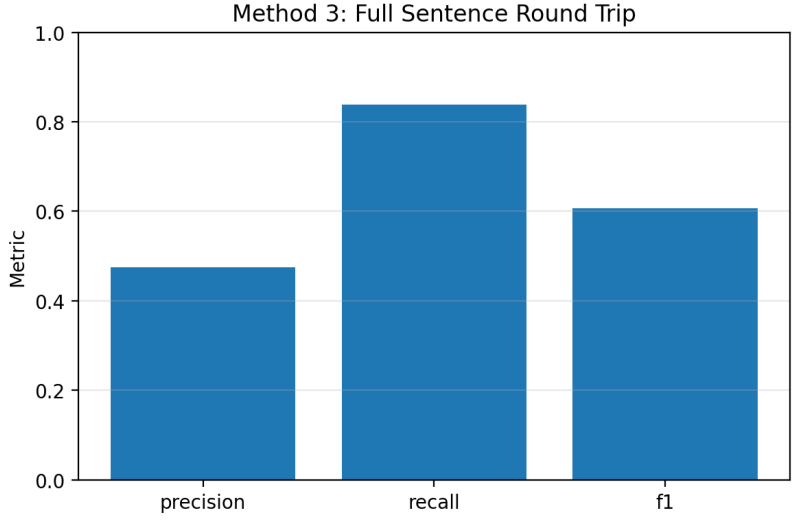
precision	recall	f1
0.476	0.926	0.629

Arabic



precision	recall	f1
0.498	0.883	0.637

Chinese



precision	recall	f1
0.475	0.838	0.607

Japanese

Spanish

precision	recall	f1
0.497	0.842	0.625

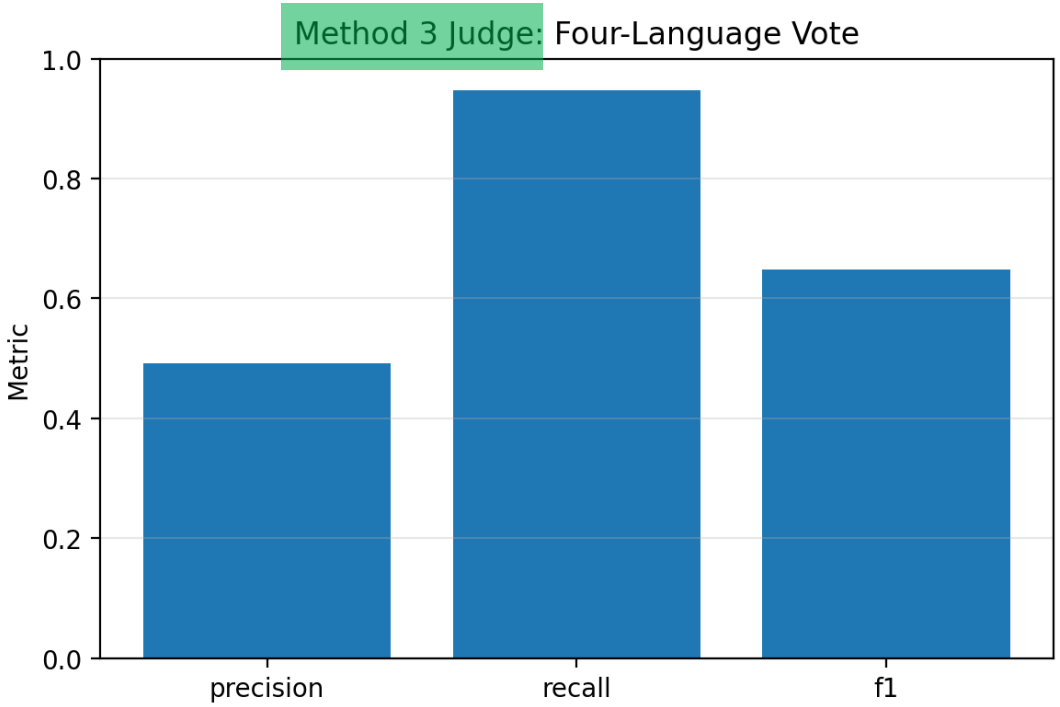
Language Performances

- Spanish closer to English in vocabulary, phrase structure, and culture
- Arabic, Chinese, and Japanese force more grammatical restructuring

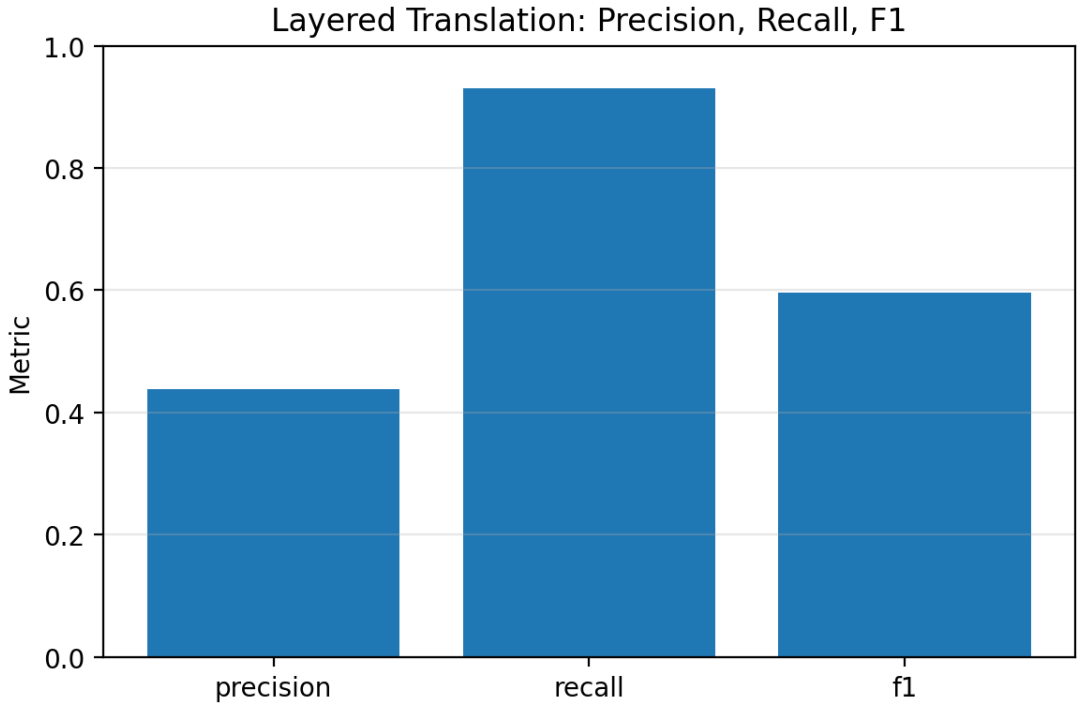
Judge System vs Layers

- Spanish (ES), Arabic (AR), Chinese (CHS), Japanese (JPN)
- **Judge**: Majority rules vote
- **Layers**: Translate English →ES →AR →CHS →JPN →English

Judge vs Layers



precision	recall	f1
0.49	0.95	0.65

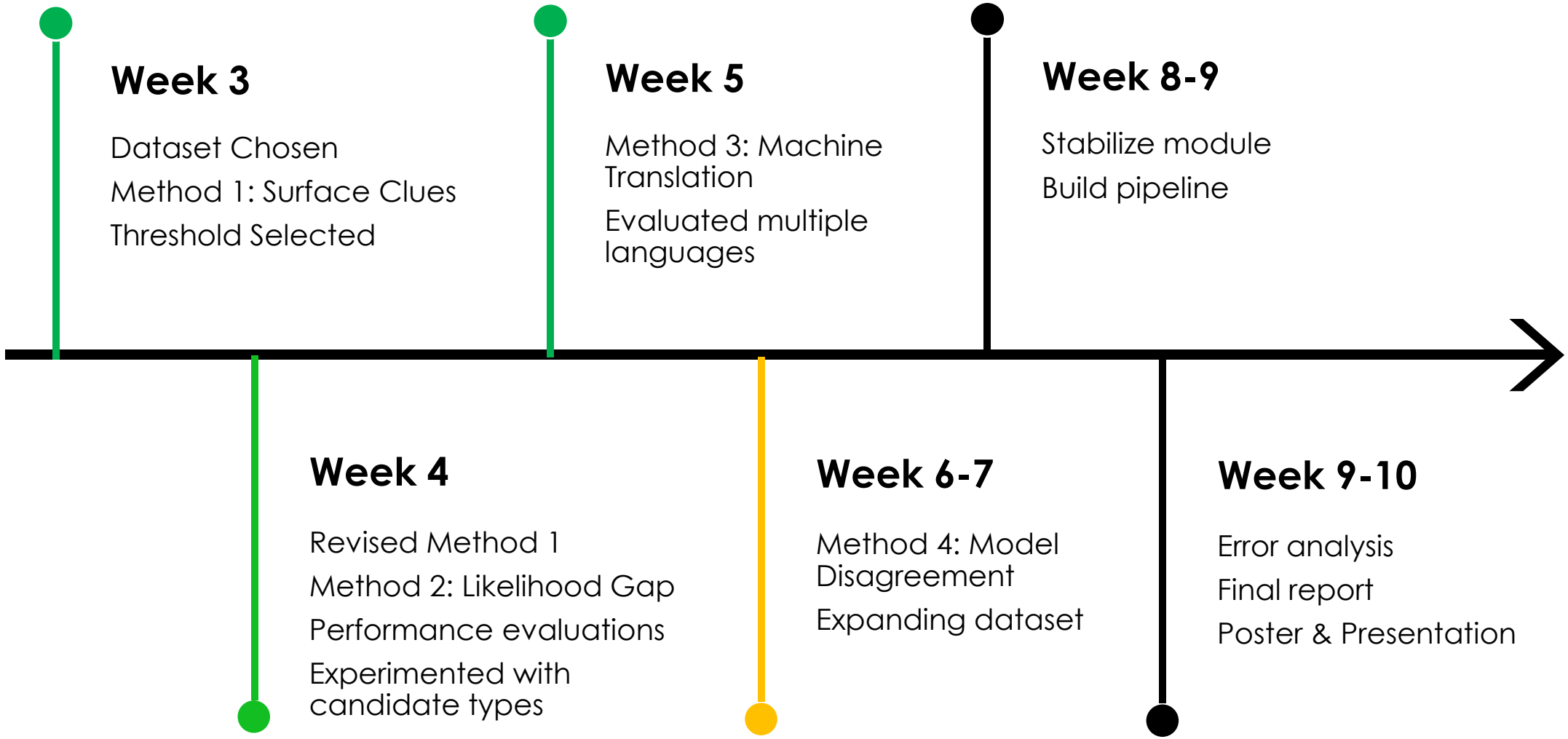


precision	recall	f1
0.438	0.931	0.595

Judge System vs Layers

- Layers accumulate in translation errors
- Majority vote rejects false positives better

Updated Timeline



Thank You

Any Questions?

By: Valerie Lopez

