WEEK 7

Detecting & Defining Specialized Language in Text

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

Valerie
Lopez

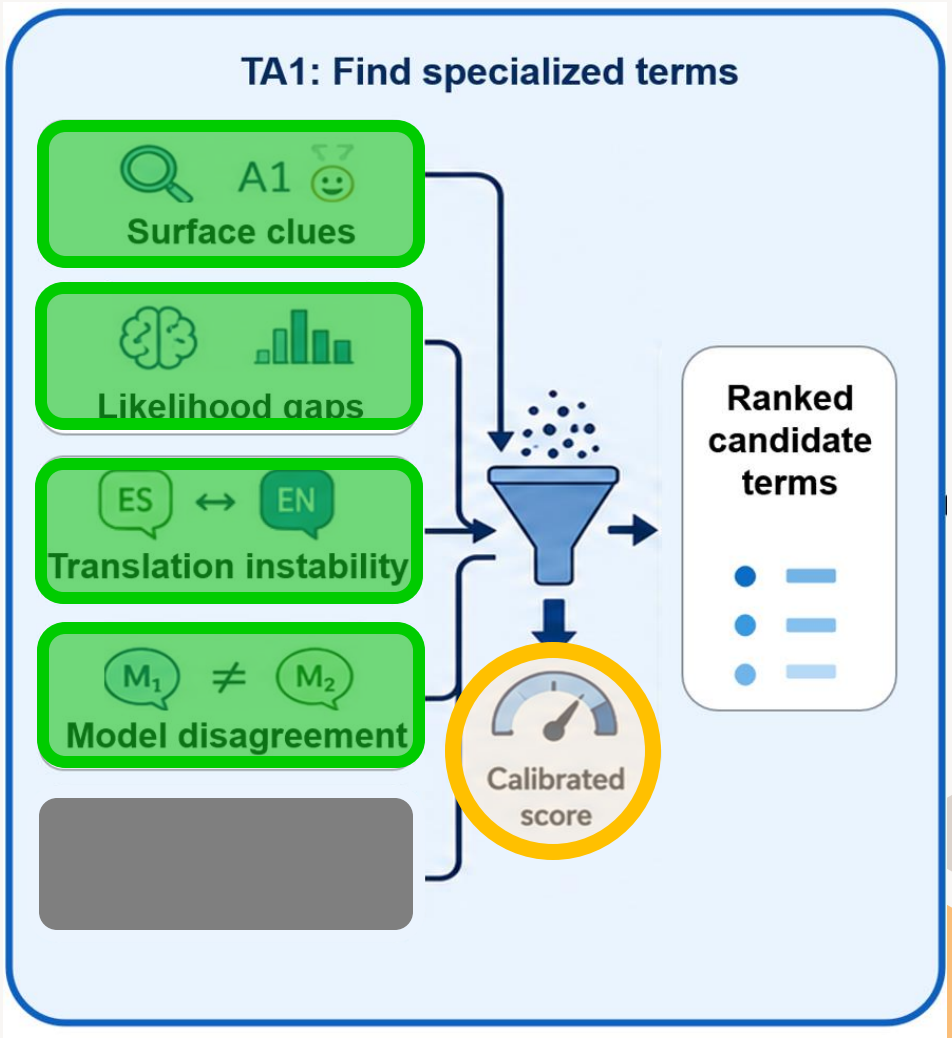


Project Recap

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

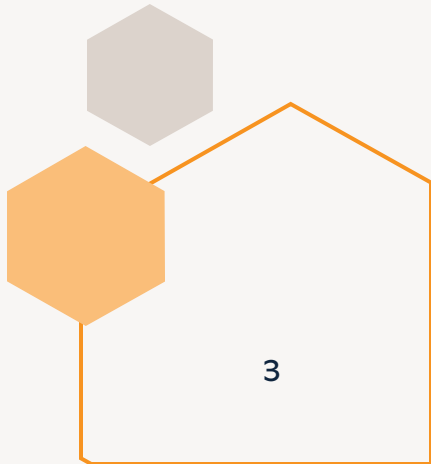
This Week:

- Compare results by candidate type
- Baseline LLM Prompting
- 2 Fusion Methods



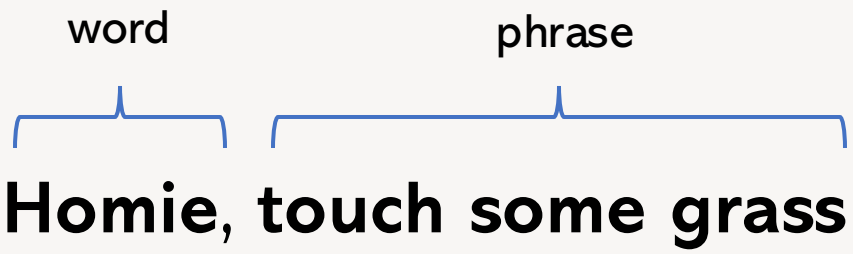
Candidate Type Recap

- Previous weeks found that phrases perform worse than words
- Best approach is to separate candidates by type and experiment with improvements for phrase search
- Method comparison results from last week were mismatched by candidate type



Candidate Type: Words

- Investigated if methods could detect words that form specialized phrases
- Version 1: Saves specialized words only, skip over phrases
- Version 2: Saves specialized words, phrases are broken down into specialized words



Version 1:
“homie”

Version 2:
“homie”
“touch”
“some”
“grass”

Word Scoring

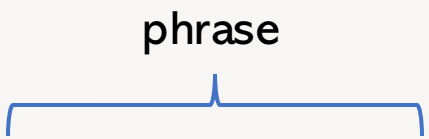
Version 1	Precision	Recall	F1
Surface Clues	**0.85**	0.28	0.42
Likelihood Gap	0.51	0.91	0.66
Machine Translation	0.56	0.98	**0.71**
Model Agreement	0.43	**0.99**	0.60

Version 2	Precision	Recall	F1
Surface Clues	**1.0**	0.24	0.38
Likelihood Gap	0.7	0.95	0.80
Machine Translation	0.77	0.97	**0.85**
Model Agreement	0.59	**0.99**	0.74

Phrase generation built off word detection is possible.

Phrase Generation

- Phrases built by permutating through sentence perform poorly
- Proposal: Take specialized predicted words, if they are consecutive, they may be a specialized phrase



Yo, the **whole squad** is **waitin'**

Predicted Specialized Words: yo, whole, squad, waitin

Consecutive Specialized Words: whole, squad

Predicted Specialized Phrase: whole squad

Phrase Generation: Permutation

- Original approach to generates phrase candidates was by permutation
- Predicted phrase candidates are compared against specialized phrase spans.

	Precision	Recall	F1
Surface Clues	0.05	0.36	0.09
Likelihood Gap	**0.06**	0.70	**0.12**
Machine Translation	0.04	**1.0**	0.07
Model Agreement	**0.06**	0.35	0.10

Consecutive Candidates

	Precision	Recall	F1
Consecutive Candidates (MT)	**0.06**	**0.83**	**0.13**
Permutation (LG)	**0.06**	0.70	0.12

- Results of predicted phrase candidates compared to specialized phrases.
- Consecutive candidates improve recall over permutations, but precision remains low
- Low precision suggests either noisy phrase labels, overly broad grouping, or both.
- Further need for synthetic data

Bro, you **ghosted** again
Dude, **touch some grass**
Yo, **whole squad** **waitin'**

Baseline Prompting Comparisons

How does our best method compare to simple LLM prompting?

	Precision	Recall	F1
Llama	0.63	0.90	0.74
Qwen	**0.70**	0.84	**0.76**
Mistral	0.56	**0.93**	0.70

Machine Translation	**0.77**	**0.97**	**0.85**

Fusion Methods

What is the best way to calibrate a single score?

✓Majority Vote

✓ Manual Weights

❑ Learned Weights

Majority Vote

	Precision	Recall	F1
1 Vote	0.58	**1.0**	0.73
2 Votes	0.68	**1.0**	0.81
3 Votes	**0.82**	0.91	**0.87**
4 Votes	1	0.24	0.38

Manual Weights

Final Score = (

0.14 * Surface Clues +

0.29 * Likelihood Gap +

0.31 * Machine Translation +

0.27 * Model Agreement

)

	F1
Surface Clues	0.38
Likelihood Gap	0.80
Machine Translation	**0.85**
Model Agreement	0.74

Weights are normalized from each method's Version 2 F1 score.

	Precision	Recall	F1
Prediction-Based	**0.82**	0.91	**0.87**
Score-Based	0.77	**0.93**	0.84

Performance Comparisons

	Precision	Recall	F1
Machine Translation (Only)	0.77	**0.97**	0.85
Qwen (LLM)	0.70	0.84	0.76
Majority Vote Fusion	**0.82**	0.91	**0.87**
Manual Weight Fusion	**0.82**	0.91	**0.87**

Phrases from Word Fusion

Does the consecutive candidates approach improve after fusing scores?

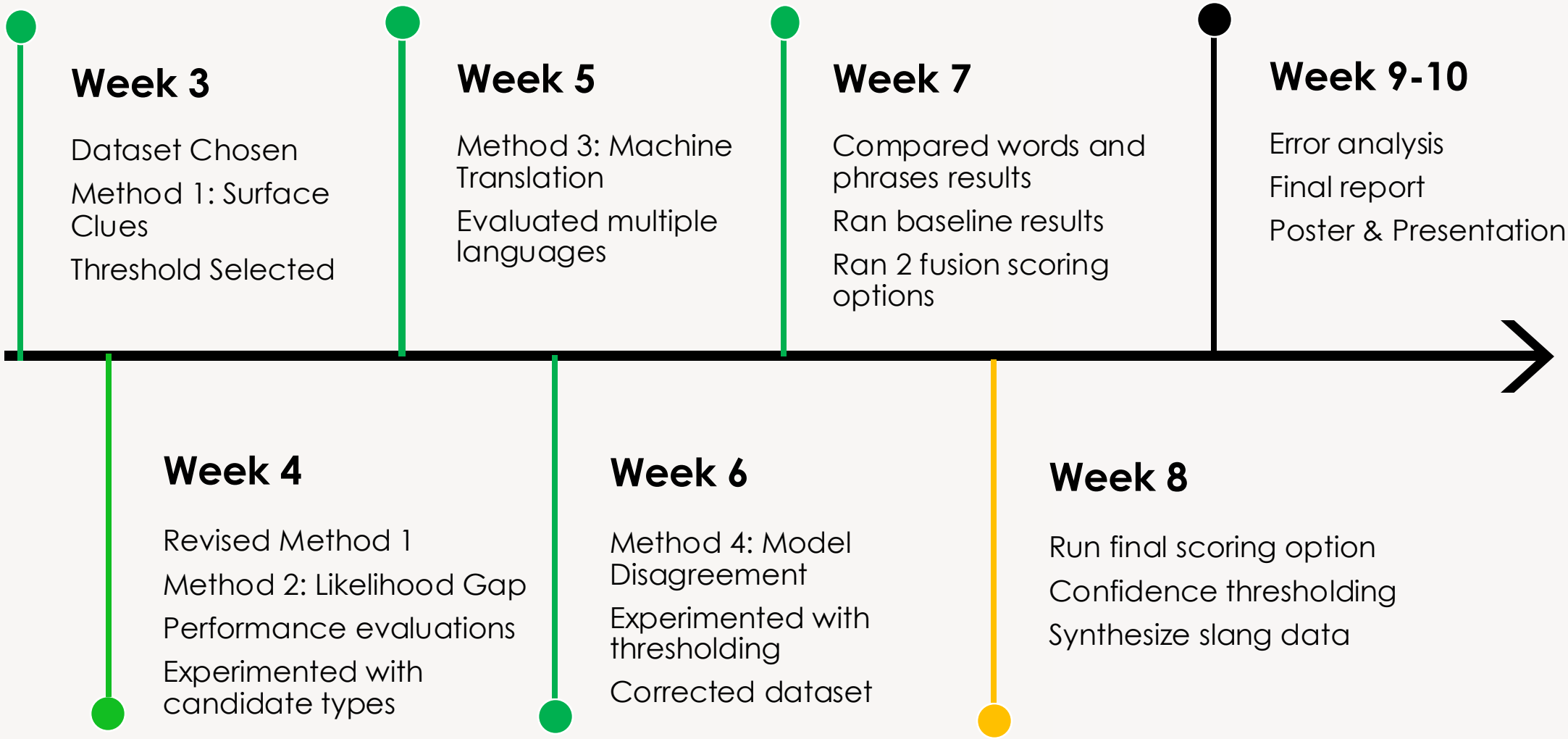
	Precision	Recall	F1
Machine Translation Only	0.06	0.83	0.13
Majority Vote	**0.08**	**0.78**	**0.15**
Manual Weight	**0.08**	**0.78**	**0.15**

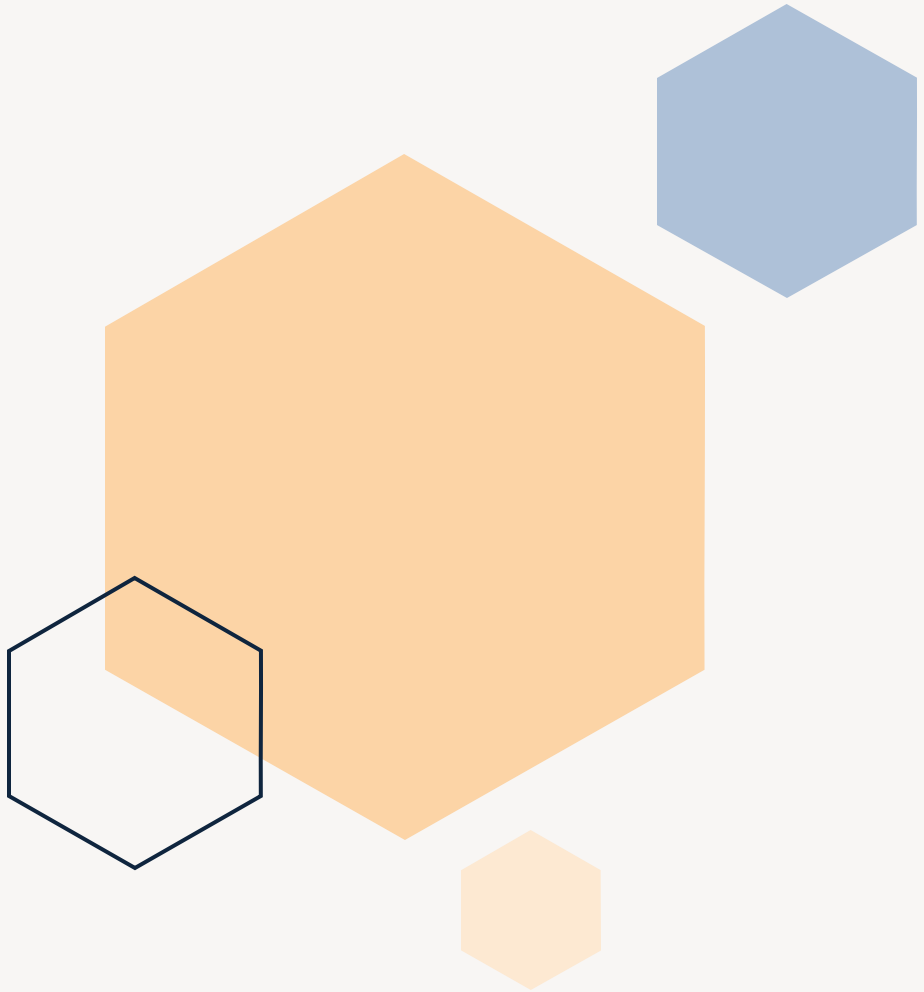
Fusion slightly improves phrase F1, but phrase recovery remains weak.

Next Steps

- ❑ Train logistic regression model to learn fusion weights
- ❑ Apply confidence thresholds
- ❑ Synthesize slang dataset

Updated Timeline





Thank You

Valerie Lopez