WEEK 8

Detecting & Defining Specialized Language in Text

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

Valerie Lopez

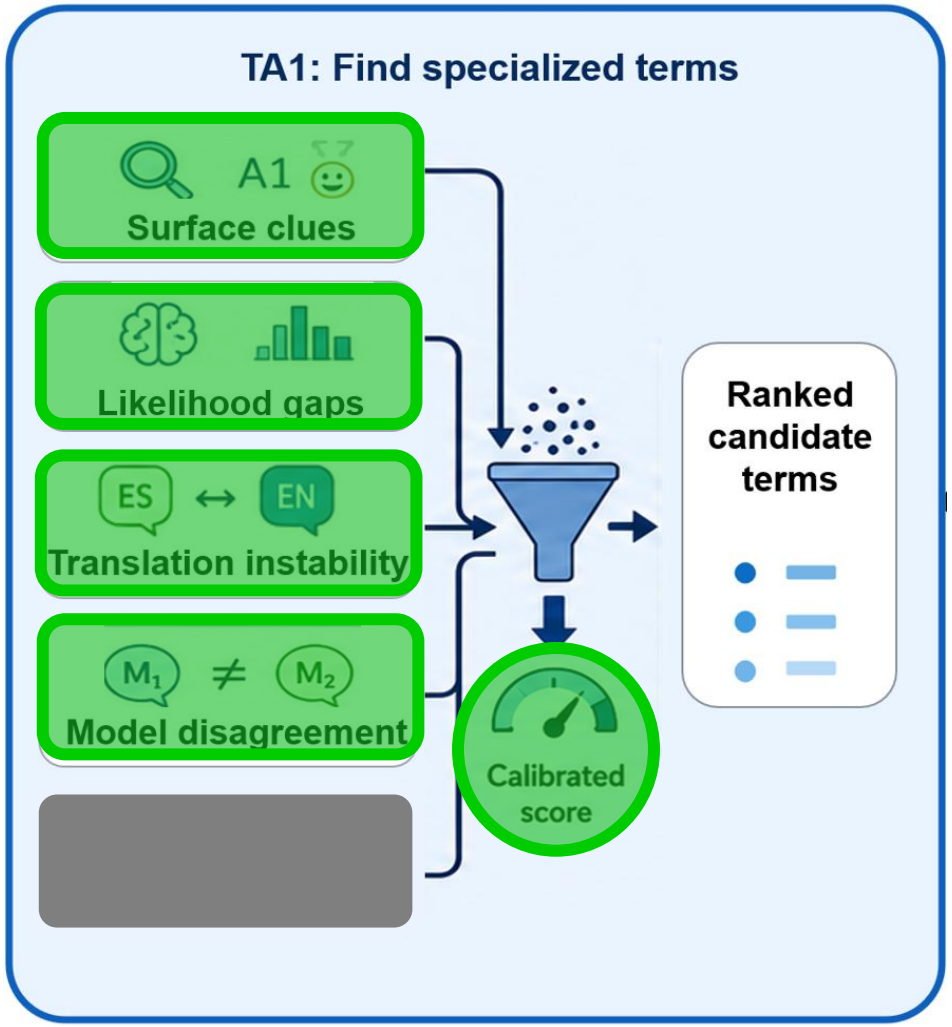


Project Recap

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

This Week:

- Built 3rd Fusion Method
- Built Pipeline Method
- Began results on new datasets



Method Overviews

Method	Overview
Surface Clues	• Run SC word detection. • Build consecutive candidates as phrases.
Likelihood Gap	• Run LG word detection. • Build consecutive candidates as phrases.
Machine Translation	• Run MT word detection. • Build consecutive candidates as phrases.
Model Agreement	• Run MA word detection. • Build consecutive candidates as phrases.

Method Overviews

Method	Overview
Majority Fusion	• Run SC, LG, MT, MA word detection. • Apply majority vote fusion. • Build consecutive candidates as phrases.
Weighted Fusion	• Run SC, LG, MT, MA word detection. • Apply manual weight fusion. • Build consecutive candidates as phrases.
Logistic Fusion	• Run SC, LG, MT, MA word detection. • Apply learned weight fusion (logistic regression model). • Build consecutive candidates as phrases.

Detection vs Fusion Methods

	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
Majority Vote	0.82	0.91	**0.87**
Manual Weights	0.82	0.91	**0.87**
Logistic Regression Weights	0.82	0.91	0.86

- Results of fusion methods on **word detection only**
- Successfully implemented logistic regression fusion
- Fusion methods are showing little improvement over using just Machine Translation

Candidate Generation Methods

- **Consecutive candidates** build specialized phrases
- **MT-first Method**
  - High recall
  - Runs on whole sentence at once which makes candidate generation more efficient
  - Other methods require us to build candidates word by word

MT-First

How it works:

1. Run Machine Translation on all word candidates
2. Use consecutive candidates approach to build phrase candidates
3. Remaining 3 methods run only on MT candidates predicted as specialized
    1. Surface Clues
    2. Likelihood Gap
    3. Model Agreement
4. Apply fusion to 3 method results

MT-First Example

Homie, touch some grass.

MT Specialized Results:

Words: homie, touch, some, grass

Phrases: touch some, some grass, touch some grass



Surface Clues

Likelihood Gap

Model Agreement

Thresholds:

MVF: 0.667

MWF: 0.5

LRF: 0.379

	SC	LG	MA	MVF	MVF P	MWF	MWF P	LRF	LRF P	is_specialized
homie	0	1	1	0.667	✓	0.8	✓	0.638	✓	✓
touch	0	1	1	0.667	x	0.8	x	0.638	x	x
some	0	1	1	0.667	x	0.8	x	0.638	x	x
grass	0	1	1	0.667	x	0.8	x	0.638	x	x
touch some	0	1	1	0.667	✓	0.8	✓	0.638	✓	x
some grass	0	0	1	0.333	x	0.4	x	0.127	x	x
touch some grass	0	1	1	0.667	✓	0.8	✓	0.638	✓	✓

MT-First Fusions

- Results of word and phrase detection
- Individual methods improved in phrase detection compared to non-pipeline results

	Precision	Recall	F1
Majority Fusion	0.658	0.815	0.728
Weighted Fusion	0.658	0.815	0.728
Logistic Fusion	0.658	0.815	0.728

Method Performance Overview

F1 scores of methods on GenZ dataset

Method	Consecutive Cand	MT-First
Surface Clues	0.376	0.190
Likelihood Gap	0.687	0.690
Machine Translation	**0.721**	-
Model Agreement	0.451	0.698
Majority Fusion	0.643	**0.728**
Weighted Fusion	0.643	**0.728**
Logistic Fusion	0.711	**0.728**

Expanding Datasets

Aim is to study how well the methods can generalize and exhibit robustness

- As we attempted to build off glossary datasets, we found they already provided sentences to test off
- Instead of synthesizing data, we apply current methods to new datasets to study behavior consistency
- New data: slang, acronyms, jargon, complex terms, idioms, and multiword expressions

Threshold experimentation:

1. Compare thresholds among datasets per method
2. High variation shows our detector needs per-dataset tuning or better calibration.
3. Create mixed set to determine best generalization

Selected Datasets

Slang

GenZ

- Baseline dataset.
- Repetitive and phrase-poor

LM-Lexicon

- Expansive slang dataset

OpenSub

- Natural slang in movie subtitles.
- **Strong benchmark candidate** because it has natural context and human annotation

Selected Datasets

Jargon

MedReadMe

- Medical readability and jargon/complex-span resource.
- **Useful for testing non-slang** specialized language.

MedJex

- Medical jargon in EHR-style text.

Acronyms

SDU

- Legal and scientific acronyms (multilingual)

MACRONYM

- Legal and scientific acronyms (multilingual)

AAAI-21

- Acronym identifier set

Selected Datasets

Multilingual

CWI-2018

- Multilingual complex word/phrase identification
- English, German, Spanish, and French
- Useful as a multilingual stress test.

PARSEME 1.3

- Verbal multiword expression dataset across many languages.

AlphaMWE

- Multilingual parallel multi word expression (MWE) data.
- **Useful for testing MT-first method** because it connects MWEs with translation behavior.

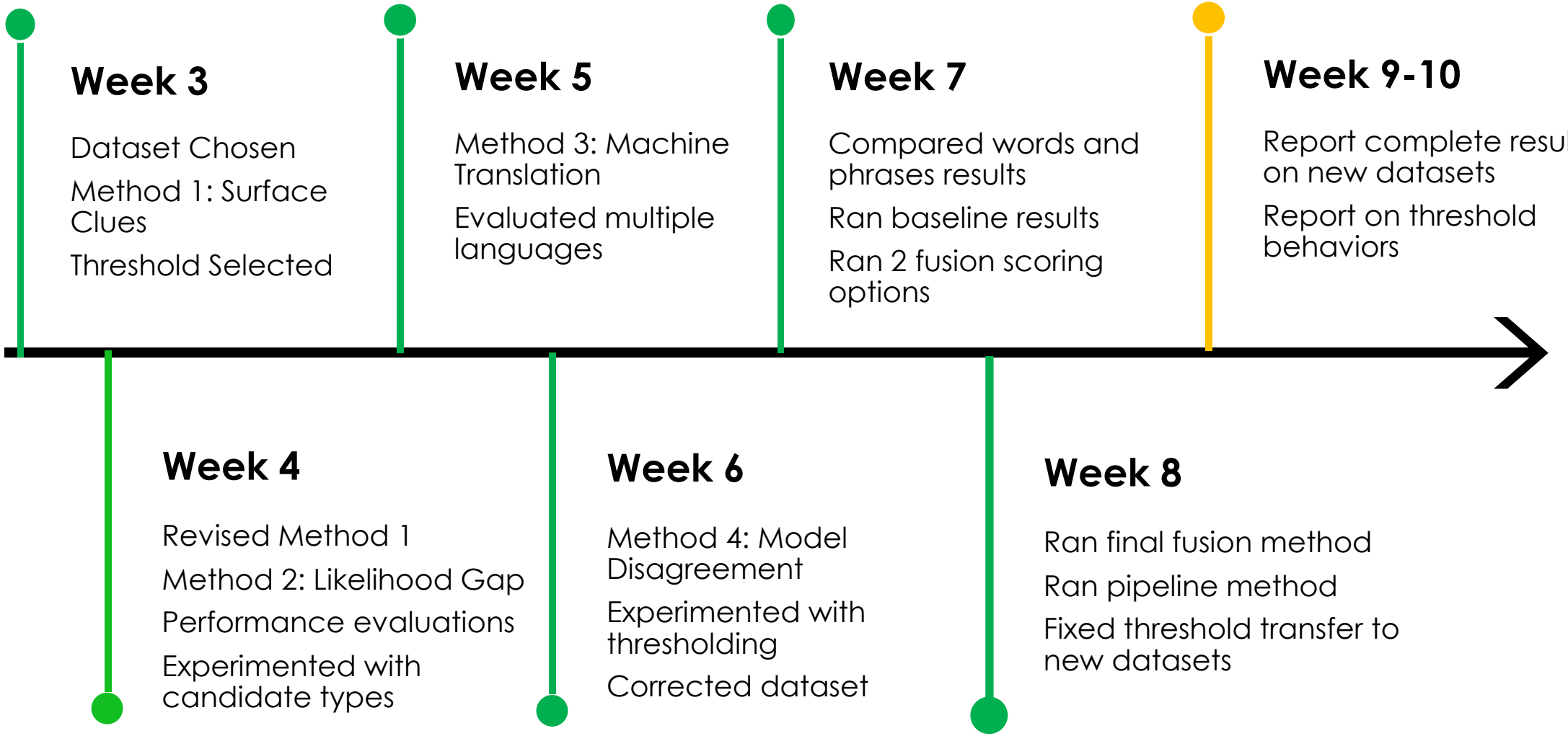
HYU-NLP/MIDAS

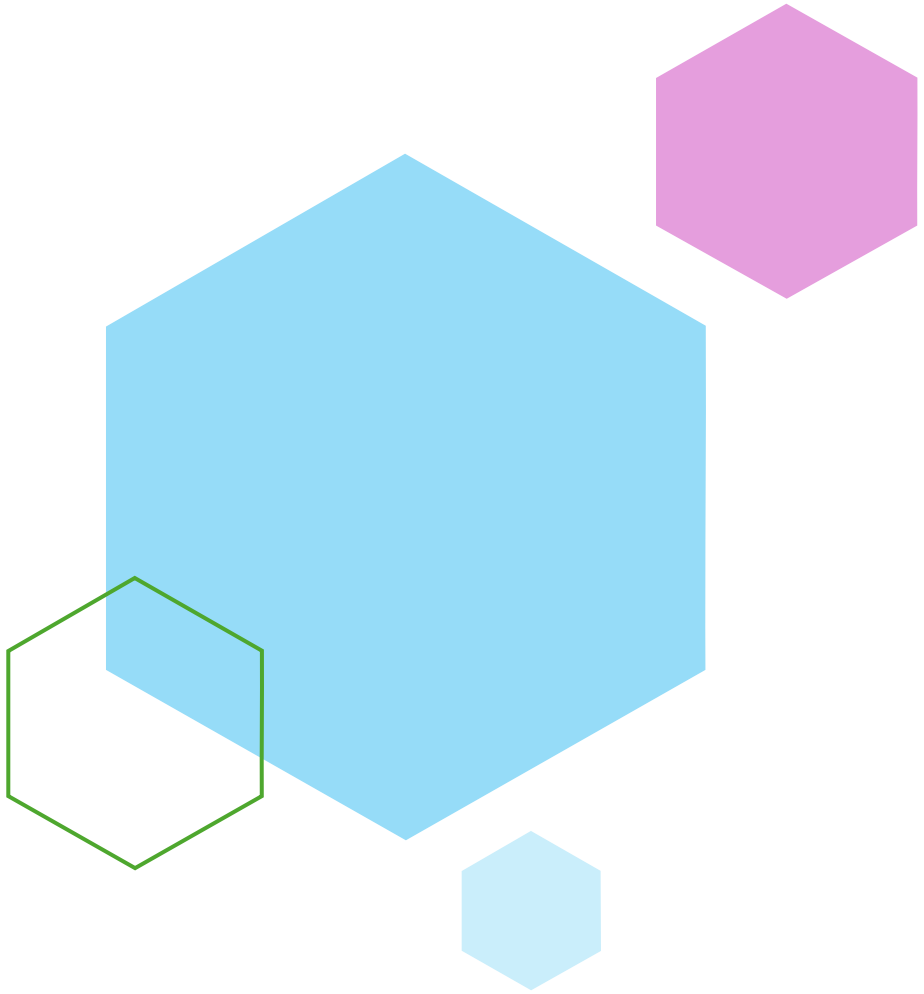
- Multilingual idiom dataset.

Next Steps

- ❑ Test on **multilingual datasets**
- ❑ Test on **medical jargon datasets**
- ❑ Test on a balanced **mixed set**
- ❑ Compare dataset best thresholds per method

Updated Timeline





Any Questions?

Valerie Lopez