INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

LTRC, IIIT Hyderabad

HashSet - A Dataset For Hashtag Segmentation

Prashant Kodali[†], **Akshala Bhatnagar**[*], **Naman Ahuja**[†]
Manish Shrivastava[†], **Ponnurangam Kumaraguru**[†]
[†]IIIT Hyderabad
prashant.kodali@research.iiit.ac.in, naman.ahuja@students.iiit.ac.in,
{m.shrivastava, pk.guru}@iiit.ac.in
[*]IIIT Delhi
akshala18012@iiitd.ac.in

Introduction

- Hashtags often encapsulate the gist, emotion, and sentiment cues and such cues are useful in downstream tasks like text classification.

- Hashtag Segmentation or Hashtag Decomposition is the task of breaking a hashtag into its constituent tokens.
  - #weareatLREC2022 --> we are at LREC2022

- Hashtags prioritize brevity and are exhibit certain quirks making the task non-trivial.

Motivation

Domain, Location, Language Specificity.

Hashtag quirks

- transliterating and mixing languages :
  - #sabkasath ---> sab ka saath
- spelling variations: #letzgooo --> #letsgo
- camel case: #WeLoveApples
- presence of special characters : #We_love_apples@1

Simple heuristics to segment like camel case, underscore

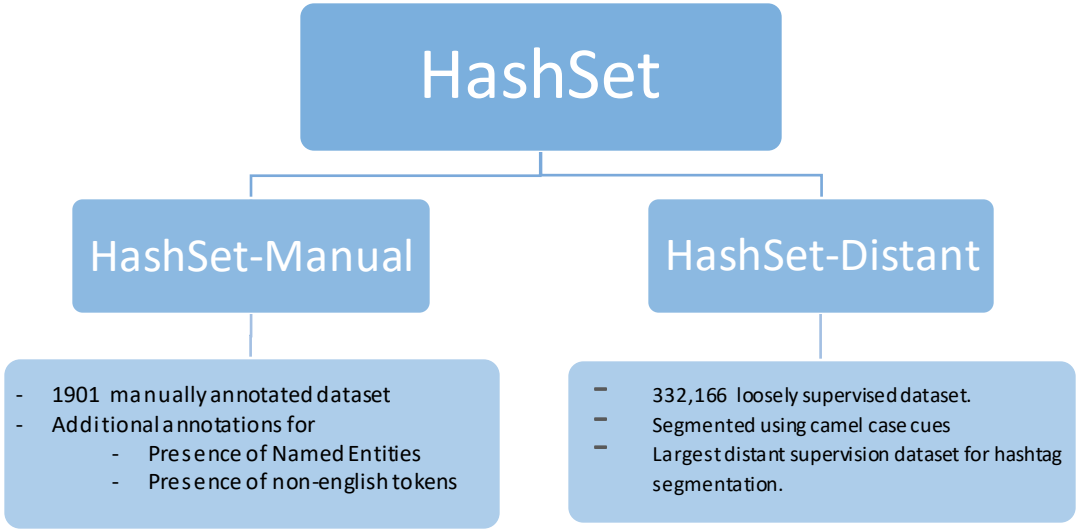
Datasets should consist of non-trivial samples.

Size and Source of Datasets

- Existing Datasets are small and extracted from a single set of tweets.

Our Dataset

- We propose **HashSet**, a dataset comprising of:



How does HashSet compare with other datasets?

Parameter	Datasets					
	STAN-Dev	STAN-Small	STAN-Large	BOUN	HashSet-Manual	HashSet-Distant
Number of Hashtags	1012	1108	11965	999	1901	332166
Avg. Hashtag Length	8.49	8.9	8.58	11.3	12.68	14.69
Avg. Number of Segments	1.75	1.78	1.74	2.37	2.49	2.8
Num of Single Token Hashtags	532 (53%)	453 (41%)	4749 (40%)	258 (26%)	396 (20.8%)	0
Num of Camel Case Hashtags	134 (13.3%)	1441 (12.0%)	108 (9.8%)	278 (27.8%)	0	332166(100%)
Num of Non-English Tokens	-	-	-	-	236 (12.4%)	-
Num of Named Entities	-	-	-	-	1414 (74.4%)	-

- Least number of single token hashtags, and no camel-cased hashtags
  - Camed cased in HashSet-Distant are also lower cased for utility in training models

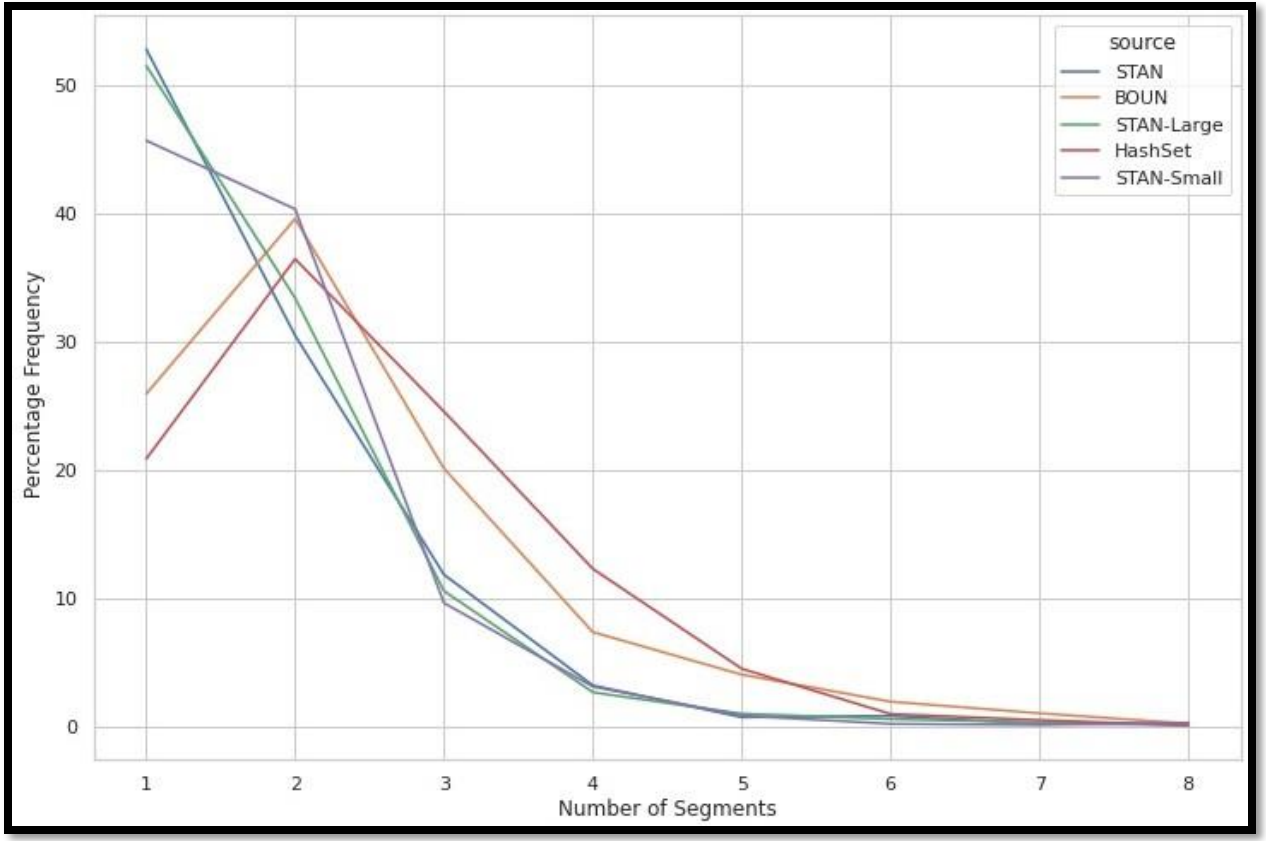
How does HashSet compare with other datasets?

Parameter	Datasets					
	STAN-Dev	STAN-Small	STAN-Large	BOUN	HashSet-Manual	HashSet-Distant
Number of Hashtags	1012	1108	11965	999	1901	332166
Avg. Hashtag Length	8.49	8.9	8.58	11.3	12.68	14.69
Avg. Number of Segments	1.75	1.78	1.74	2.37	2.49	2.8
Num of Single Token Hashtags	532 (53%)	453 (41%)	4749 (40%)	258 (26%)	396 (20.8%)	0
Num of Camel Case Hashtags	134 (13.3%)	1441 (12.0%)	108 (9.8%)	278 (27.8%)	0	332166(100%)
Num of Non-English Tokens	-	-	-	-	236 (12.4%)	-
Num of Named Entities	-	-	-	-	1414 (74.4%)	-

- Annotations for Names Entities, Non-english tokens

How does HashSet compare with other datasets?

- Higher Hashtag length
- Higher average number of tokens.



Dataset Collection and Annotation

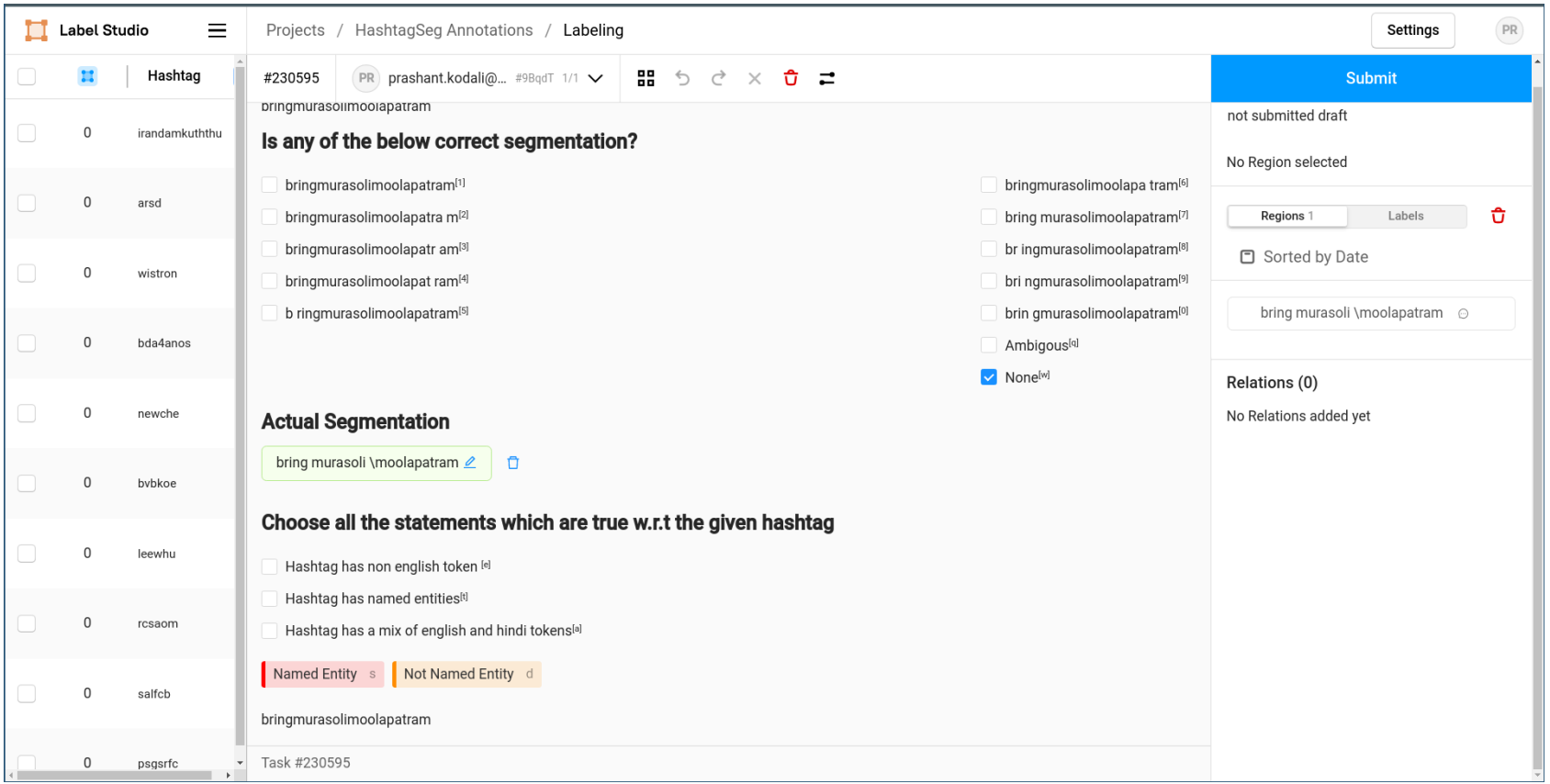
- To create a large set of hashtags, we used Twitter API
  - trending hashtags across different locations for the period May-Oct 2021
  - hashtags from a collection of tweets for trending hashtags during for April – May 2019.

- Total collected Hashtags : 841,520
  - Roman script: 731,357 hashtags
  - Camel case: 319,497 hashtags

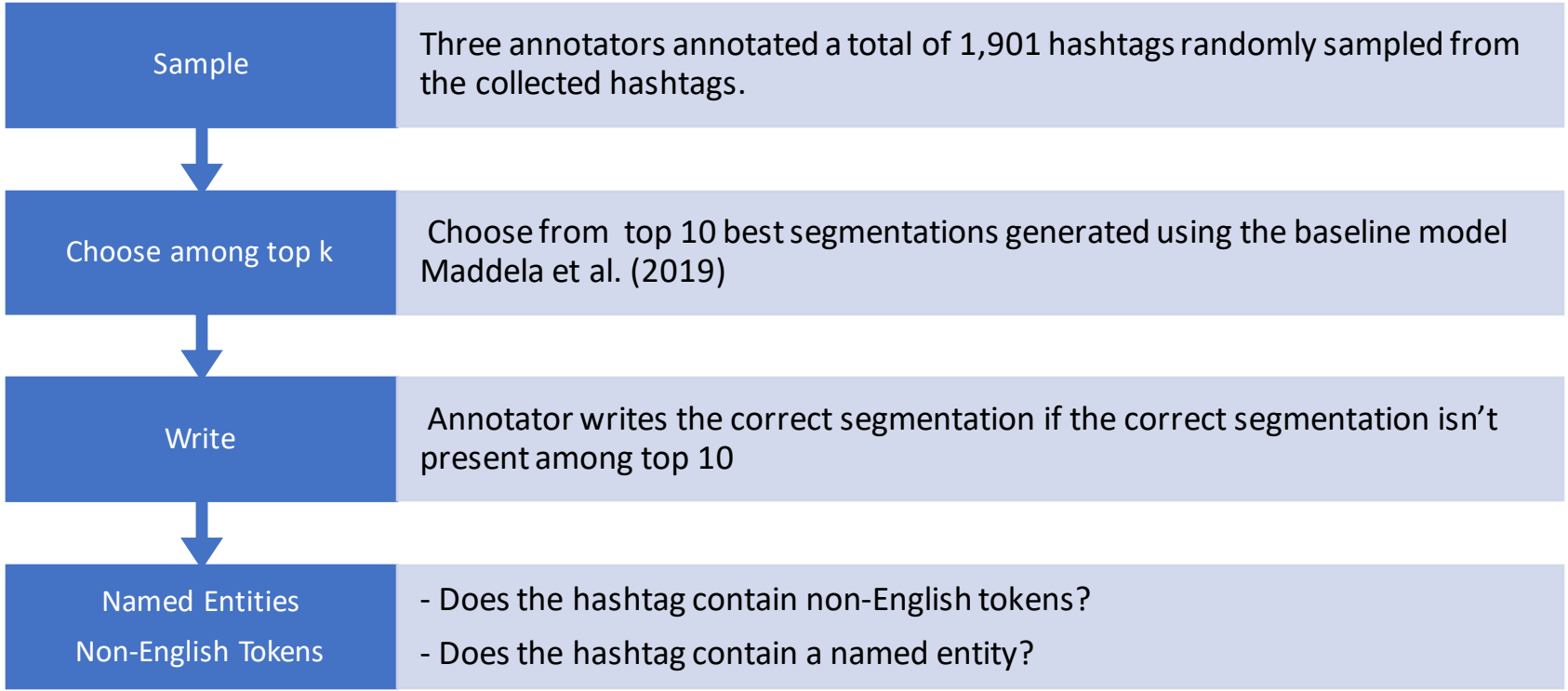
- We randomly sample hashtags from the collected hashtags, and three annotators carry out the annotations.

- We use LabelStudio tool for annotations.

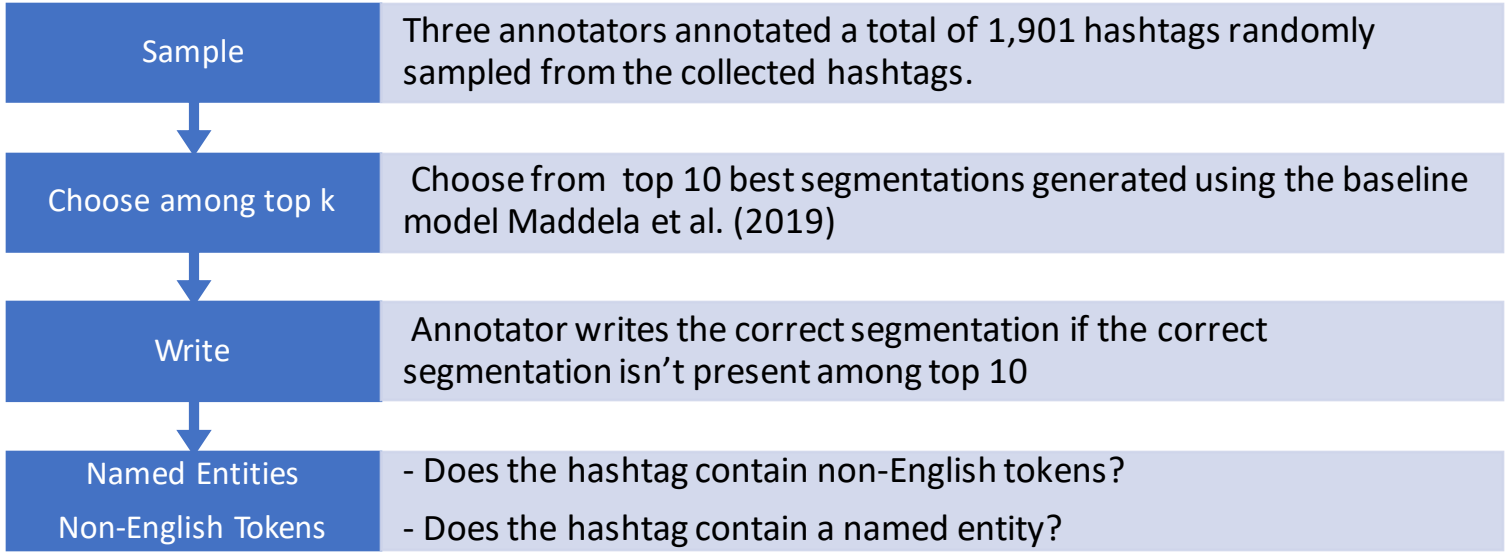
Annotation Process - Manual



Annotation Process - Manual



Annotation Process - Manual

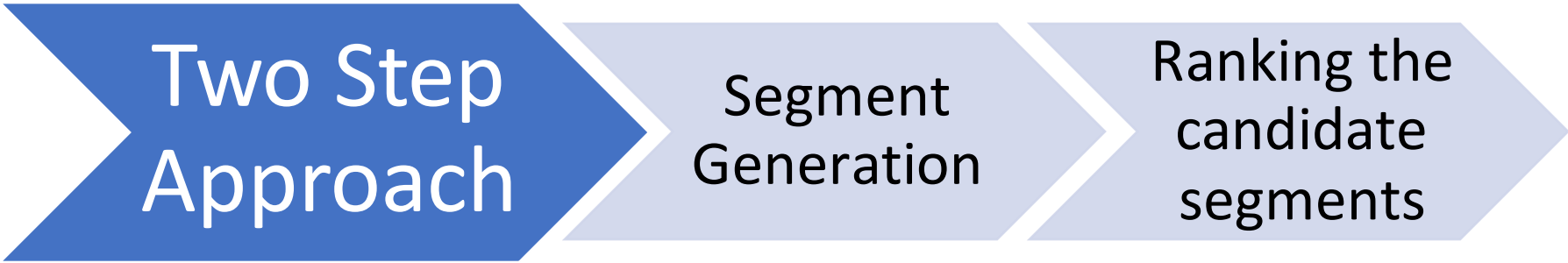


- We use model predictions to speed up the annotation process
- Ambiguous Hashtags:
  - Hashtags that the annotators are unable to segment with certainty and mark as ambiguous;
  - 89 such hashtags were found in the annotation process and were excluded.

Annotation Process - Distant

- Motivation : Manual annotation of the hashtags is a time-consuming process.
- Nearly 43% of the collected hashtags are written in camel case, and 3% of hashtags have underscores.
- Regular expression-based patterns to split the camel case hashtags.
- Store lower cased version of hashtags and their segments
- Cases where such heuristic would fail
  - #CostofViraatvacation --> Costof Viraatvacation
    - whereas the correct segmentation would be Cost of Viraat vacation

Baseline Models – Common Architecture



Baseline Models

Multitask Pairwise Neural Ranking (MPNR)

Maddela et al. (2019)

- Candidate Generation: word breaker + statistical LMs
- Candidates are reranked using neural architectures.

Hashformer

Rodrigues et al. (2021)

- Ensemble of transformer-based language models to generate segments, and re-ranking
- Zero-shot architecture

Experimental Setup

- We use the publicly available implementation of the SOTA models and reproduce the results on all datasets for further analysis.
- For MPNR model, we reproduce the results using the language models released by authors, and for Hashformers we use the publicly available versions of GPT-2, BERT.
- For each dataset, we generate top-10 segmentations. We evaluate the models using accuracy measure.
- A sample is classified as correct if the correct segmentation figures in top-n segments, where n ranges from 1 to 10.

Model Performance

Architecture	Dataset	Accuracy @ top -n					
		n=1	n=2	n=5	n=7	n=9	n=10
MPNR	BOUN	81.6	88.09	90.29	90.69	90.69	90.69
	STAN-Dev	73.12	78.16	81.92	82.71	82.81	82.81
	STAN-Small	82.76	86.19	86.82	86.82	86.82	86.82
	STAN-Large	63.78	73.10	74.73	74.75	74.75	74.75
	HashSet-Manual	41.93	45.98	47.5	47.71	47.71	47.71
Hashformers	BOUN	83.68	87.69	91.39	99.00	99.30	99.30
	STAN-Dev	80.04	84.49	90.02	98.72	99.51	99.60
	STAN-Small	80.05	85.11	88.90	97.11	97.38	97.38
	STAN-Large	72.17	75.74	79.25	85.38	85.82	85.86
	HashSet-Manual	56.71	68.54	78.22	91.53	94.00	94.37

- MPNR and Hashformer, perform well for BOUN, $STAN_{dev}$, $STAN_{small}$, $STAN_{large}$.

Model Performance

Architecture	Dataset	Accuracy @ top -n					
		n=1	n=2	n=5	n=7	n=9	n=10
MPNR	BOUN	81.6	88.09	90.29	90.69	90.69	90.69
	STAN-Dev	73.12	78.16	81.92	82.71	82.81	82.81
	STAN-Small	82.76	86.19	86.82	86.82	86.82	86.82
	STAN-Large	63.78	73.10	74.73	74.75	74.75	74.75
	HashSet-Manual	41.93	45.98	47.5	47.71	47.71	47.71
Hashformers	BOUN	83.68	87.69	91.39	99.00	99.30	99.30
	STAN-Dev	80.04	84.49	90.02	98.72	99.51	99.60
	STAN-Small	80.05	85.11	88.90	97.11	97.38	97.38
	STAN-Large	72.17	75.74	79.25	85.38	85.82	85.86
	HashSet-Manual	56.71	68.54	78.22	91.53	94.00	94.37

- Hashformer consistently outperforms MPNR across datasets. Accuracies improve as n approaches 10.

Model Performance

Architecture	Dataset	Accuracy @ top -n					
		n=1	n=2	n=5	n=7	n=9	n=10
MPNR	BOUN	81.6	88.09	90.29	90.69	90.69	90.69
	STAN-Dev	73.12	78.16	81.92	82.71	82.81	82.81
	STAN-Small	82.76	86.19	86.82	86.82	86.82	86.82
	STAN-Large	63.78	73.10	74.73	74.75	74.75	74.75
	HashSet-Manual	41.93	45.98	47.5	47.71	47.71	47.71
Hashformers	BOUN	83.68	87.69	91.39	99.00	99.30	99.30
	STAN-Dev	80.04	84.49	90.02	98.72	99.51	99.60
	STAN-Small	80.05	85.11	88.90	97.11	97.38	97.38
	STAN-Large	72.17	75.74	79.25	85.38	85.82	85.86
	HashSet-Manual	56.71	68.54	78.22	91.53	94.00	94.37

- On HashSet-Manual dataset, performance of both models degrades.
- Degradation in MPNR is much starker compared to Hashformer.

Model Performance

Architecture	Dataset	Accuracy @ top -n					
		n=1	n=2	n=5	n=7	n=9	n=10
MPNR	BOUN	81.6	88.09	90.29	90.69	90.69	90.69
	STAN-Dev	73.12	78.16	81.92	82.71	82.81	82.81
	STAN-Small	82.76	86.19	86.82	86.82	86.82	86.82
	STAN-Large	63.78	73.10	74.73	74.75	74.75	74.75
	HashSet-Manual	41.93	45.98	47.5	47.71	47.71	47.71
Hashformers	BOUN	83.68	87.69	91.39	99.00	99.30	99.30
	STAN-Dev	80.04	84.49	90.02	98.72	99.51	99.60
	STAN-Small	80.05	85.11	88.90	97.11	97.38	97.38
	STAN-Large	72.17	75.74	79.25	85.38	85.82	85.86
	HashSet-Manual	56.71	68.54	78.22	91.53	94.00	94.37

- Drop in performance can be attributed to
  - MPNR relies on statistical LMs, which have lower coverage
  - Large domain shift

Model Performance

Architecture	% containing named entities	% containing non-English tokens
MPNR	77.17	17.61
Hashformer	77.57	33.64

- Camel cased Hashtags Segmentation Accuracy for n = 1
  - MPNR ：17.8% incorrectly segmented
  - Hashformer ：15.8% incorrectly segmented

Model Performance

Architecture	% containing named entities	% containing non-English tokens
MPNR	77.17	17.61
Hashformer	77.57	33.64

- Lower error rate in camel cased hashtags → camel case points a strong signal for segmentation
- Validates our hypothesis that camel cased hashtags are relatively easy for models to segment.

Limitations

- Collection of Hashtags. Hashtag collection period was limited to window of 6 months.

- Domain/region specificity:
  - Hashtags collection is sourced from Indian cities and collection of Indian election,
  - Named entities are of Indian origin,
  - Non-English tokens belong to Indian languages, with the majority being romanized Hindi tokens.

Future Directions

- Named entity recognizer that works on an unsegmented text – A strong indicator of segmentation.

- Demonstrating Utility of distant supervised data on model training and performance

Thank You!