

Unsupervised Tokenization Learning

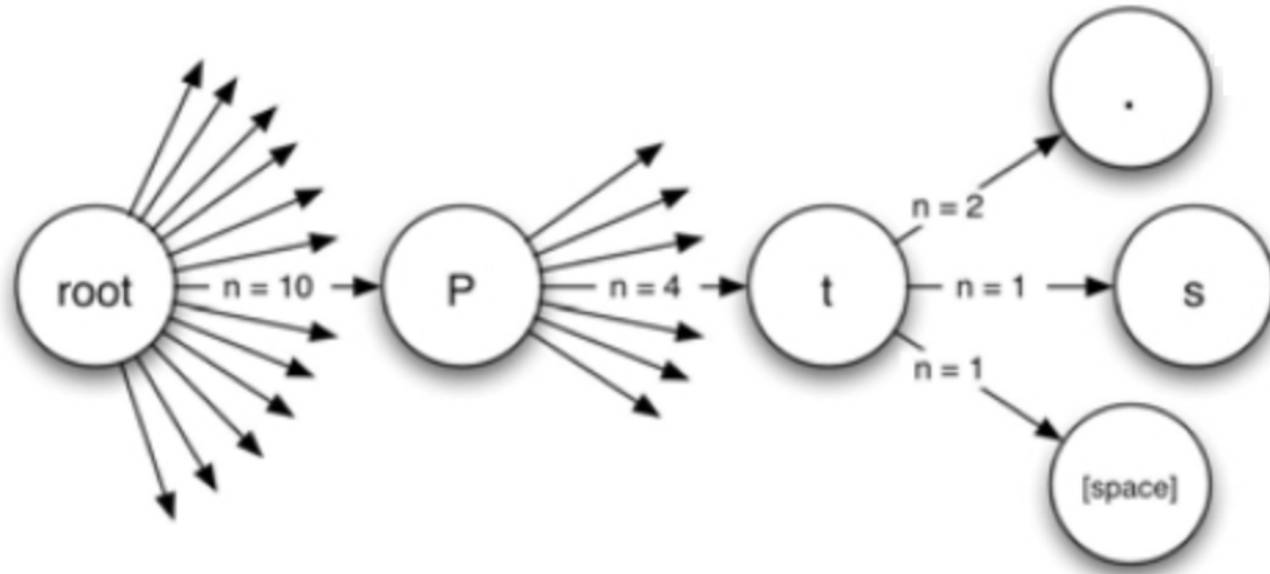
Anton Kolonin
Aigents,
SingularityNET Foundation,
Novosibirsk State University
akolonin@gmail.com



Vignav Ramesh
Harvard University
SingularityNET Foundation
vignavramesh@college.harvard.edu



Background – Tokenization as Language Modeling



Metrics/Indicators:

Conditional Probability
“Transition Freedom”

Trie data structure. The probability of observing an ‘s’ given the preceding string “Pt” is $\frac{1}{4}$, or 25%. The freedom following “pt” is 3.

Copyright ©2007 AMIA - All rights reserved. Jesse O. Wrenn, Peter D. Stetson, and Stephen B. Johnson. 2007. An unsupervised machine learning approach to segmentation of clinician-entered free text. AMIA Annu Symp Proc. 2007; 2007: 811–815.

Unsupervised Text Segmentation (Tokenization)

Metrics/Indicators:

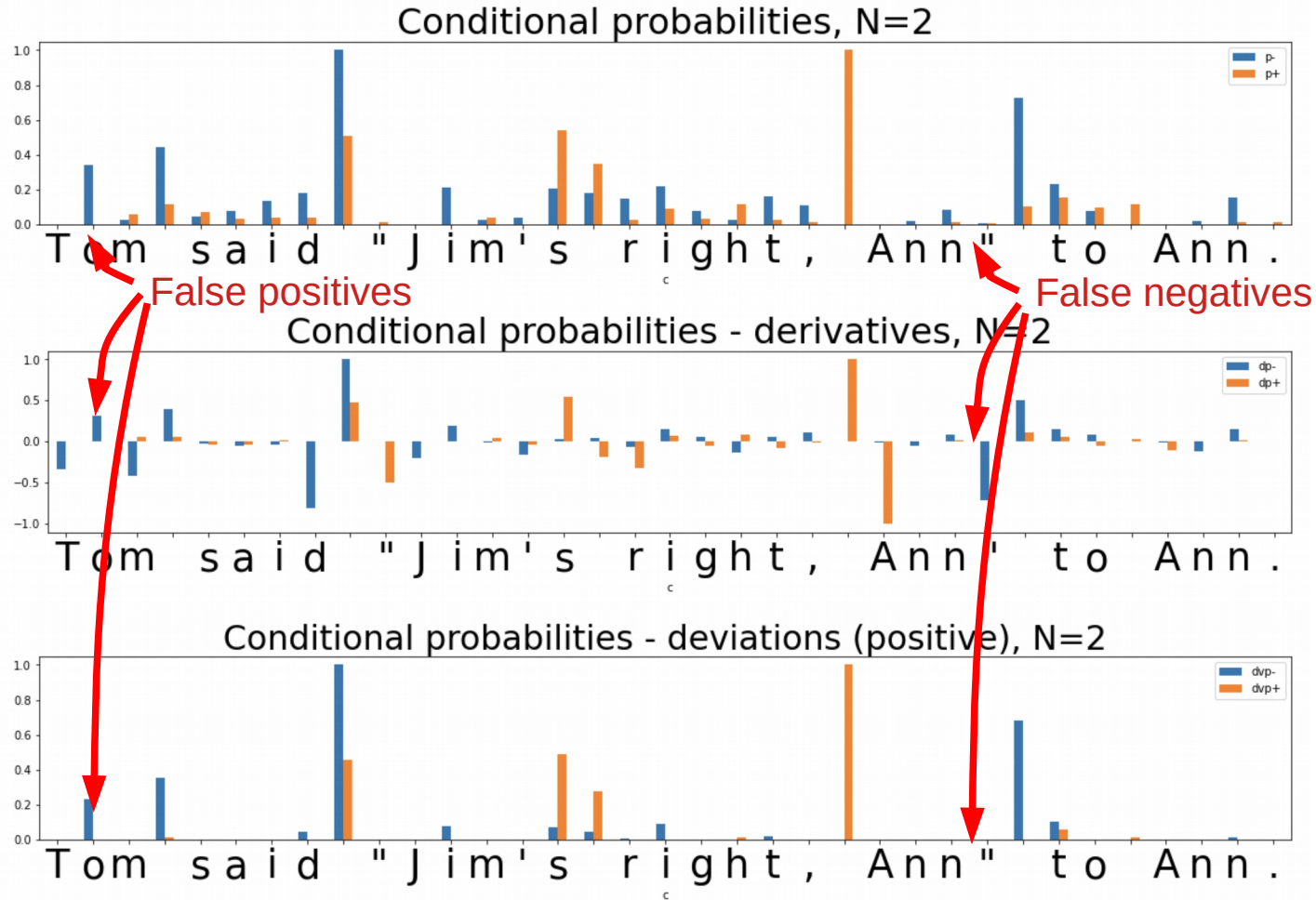
Ngram (Character)

Conditional
Probability

(of Transition)

$$P(\text{Ngram}_{n+1})/P(\text{Ngram}_n)$$

$$P(\text{"m "})/P(\text{"m"})$$

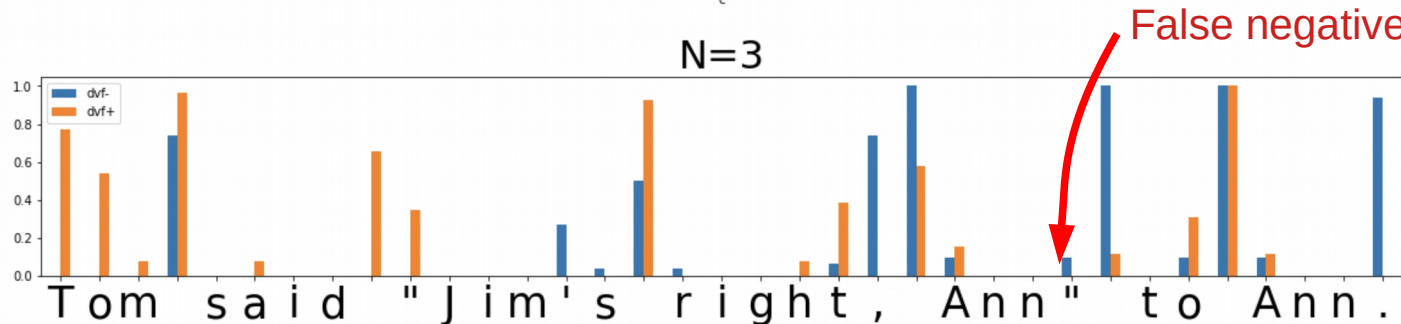
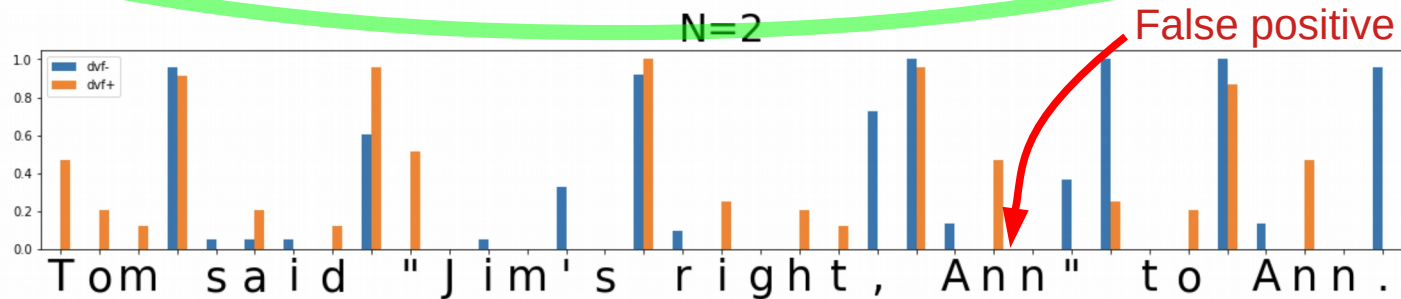
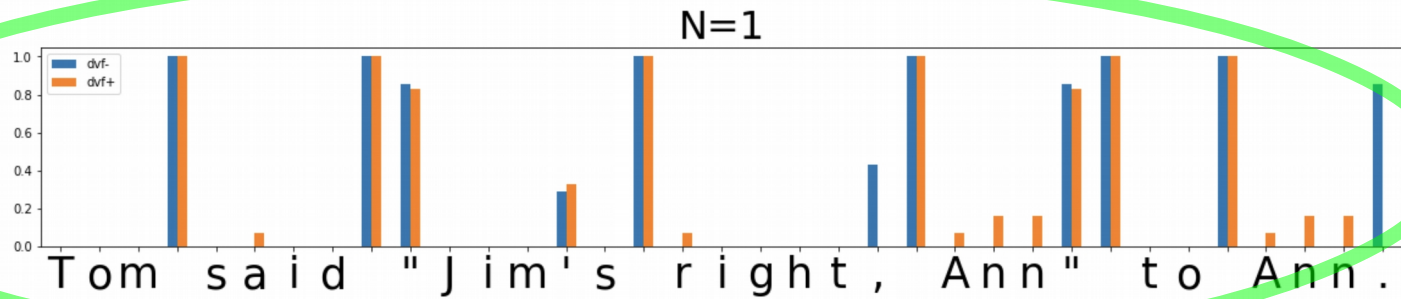


Unsupervised Text Segmentation (Tokenization)

**Metrics/
Indicators:**

Transition
Freedom
Deviation

(varying “N”
of N-gram)



Unsupervised Text Segmentation (Tokenization)

English

Hyper-Parameters:

Metric:
Transition
Freedom

Threshold
for model
compression

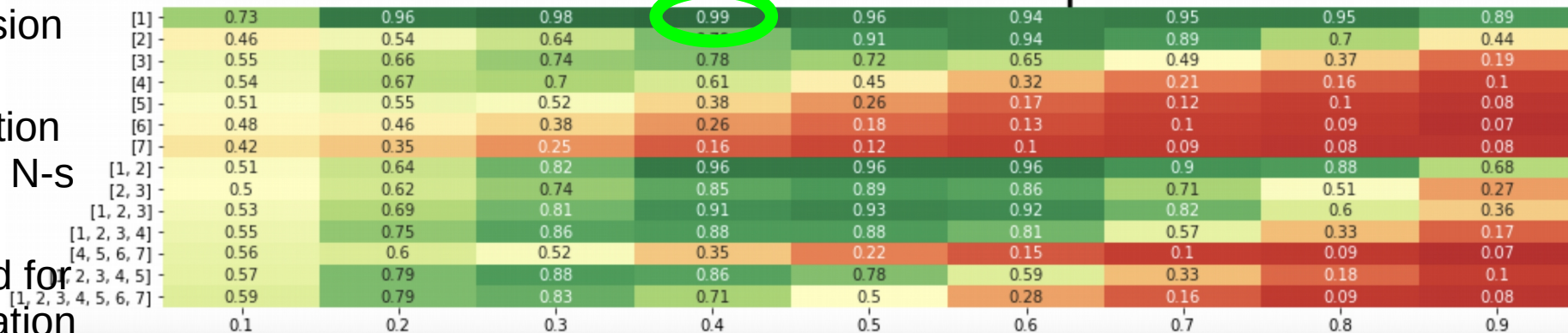
Combination
of Ngram N-s

Threshold for
segmentation

F1 - Brown ddf- & ddf+ filter=0 parameters=10967135



F1 - Brown ddf- & ddf+ filter=0.0001 parameters=8643703



Results – Freedom-based Tokenization against Lexicon-based one (referring to Rule-based)

Language	Tokenizer	Tokenization F_1	Lexicon Discovery Precision
English	Freedom-based	0.99	0.99 (vs. 1.0)
English	Lexicon-based*	0.99	-
Russian	Freedom-based	1.0	1.0 (vs. 1.0)
Russian	Lexicon-based*	0.94	-
Chinese	Freedom-based	0.71	0.92 (vs. 0.94)
Chinese	Lexicon-based*	0.83	-

**Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency*

Preprint: <https://arxiv.org/abs/2205.11443>