Unsupervised Tokenization Learning

Anton Kolonin, Aigents, SingularityNET Foundation, Novosibirsk State University
akolonin@gmail.com
Vignav Ramesh, Harvard University, SingularityNET Foundation
vignavramesh@college.harvard.edu

Abstract: In the presented study, we discover that the so-called “transition freedom” metric appears superior for unsupervised tokenization purposes in comparison to statistical metrics such as mutual information and conditional probability, providing F-measure scores in range from 0.71 to 1.0 across explored multilingual corpora. We find that different languages require different offshoots of that metric (such as derivative, deviation, and “peak values”) for successful tokenization. Larger training corpora do not necessarily result in better tokenization quality, while compressing the models by eliminating statistically weak evidence tends to improve performance. The proposed unsupervised tokenization technique provides quality better than or comparable to lexicon-based ones, depending on the language.

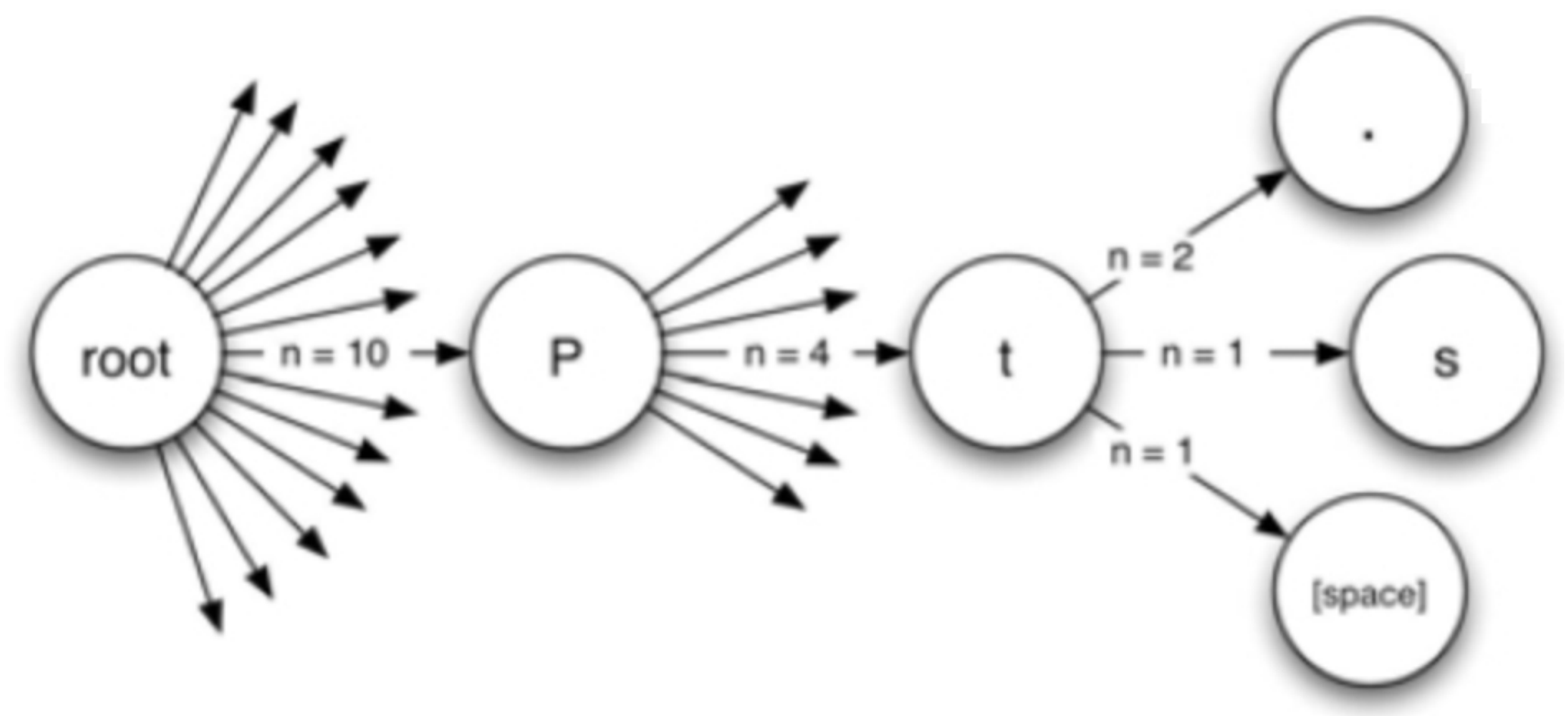

Figure 1: Background: Trie data structure. The probability of observing an ‘s’ given the preceding string “Pt” is ¼, or 25%. The freedom following “pt” is 3.
Copyright ©2007 AMIA - All rights reserved. Jesse O. Wrenn, Peter D. Stetson, and Stephen B. Johnson. 2007. An unsupervised machine learning approach to segmentation of clinician-entered free text. AMIA Annu Symp Proc. 2007; 2007: 811–815.

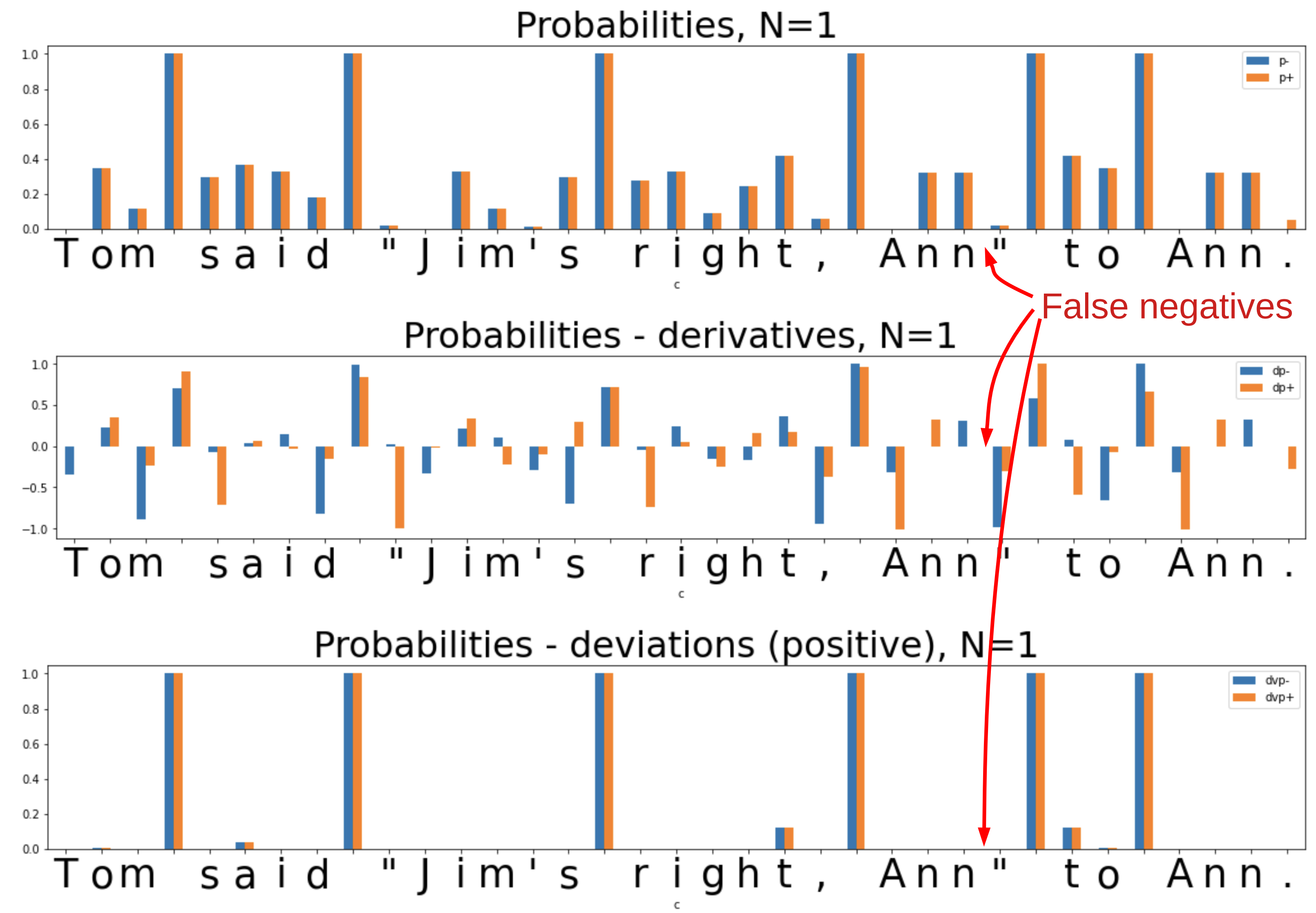

Figure 2: Using probabilities (p) and derived metrics such as derivative (dp) and deviation (dvp) in forward (dp+, dvp+) and backward (dp-, dvp-) traversals. It is clearly seen that punctuation marks cannot be isolated from words. Orange bars correspond to metrics evaluated for forward traversal, blue bars – to backward traversal.

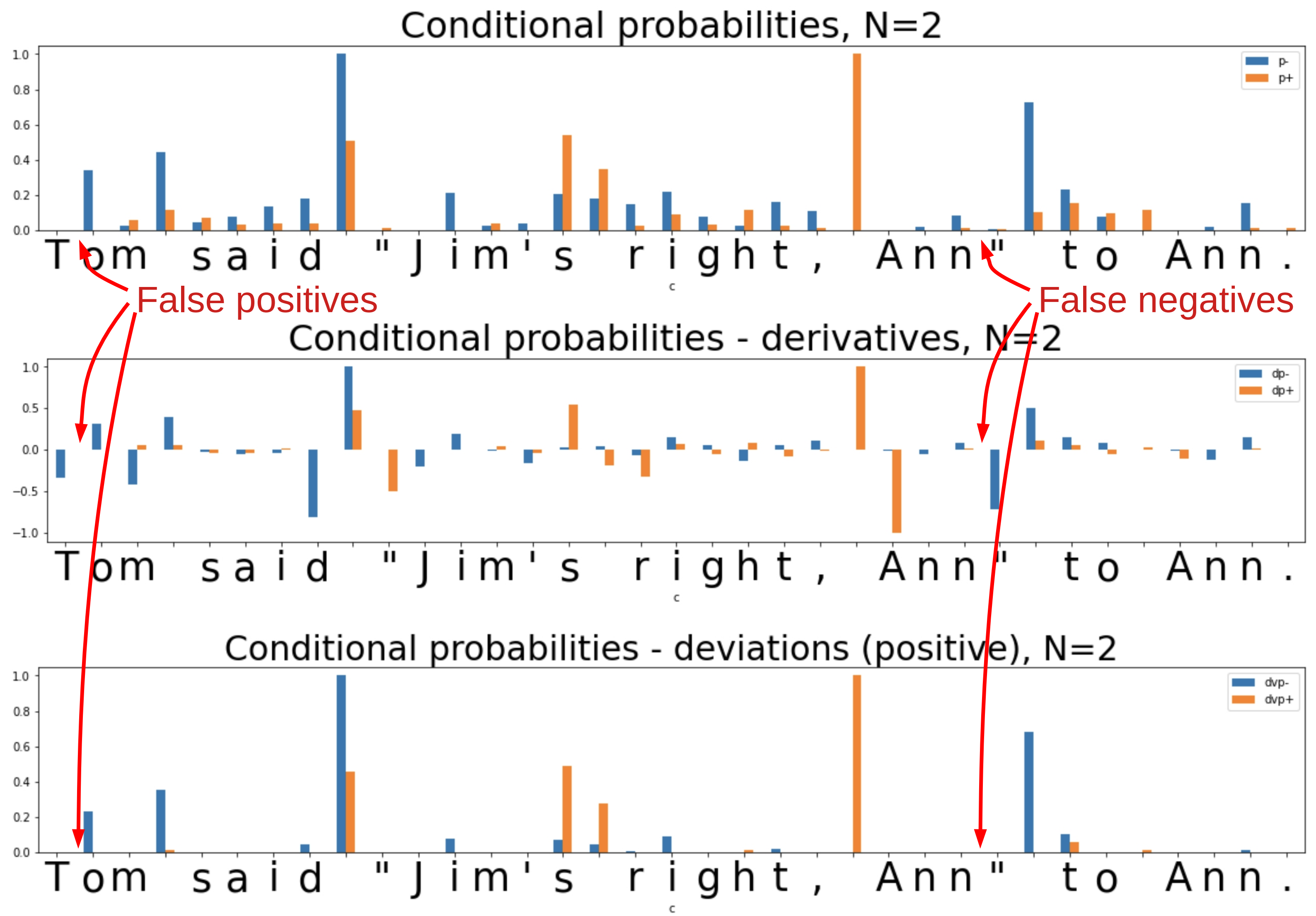

Figure 3: Using conditional probabilities (p) and derived metrics such as derivative in forward (dp+) and backward (dp-) transitions and deviation (dvp+ and dvp-, respectively) computed on N-grams with N=2. It is clearly seen that punctuation marks cannot be isolated from words, and some of the words are mistakenly disassembled into pieces.

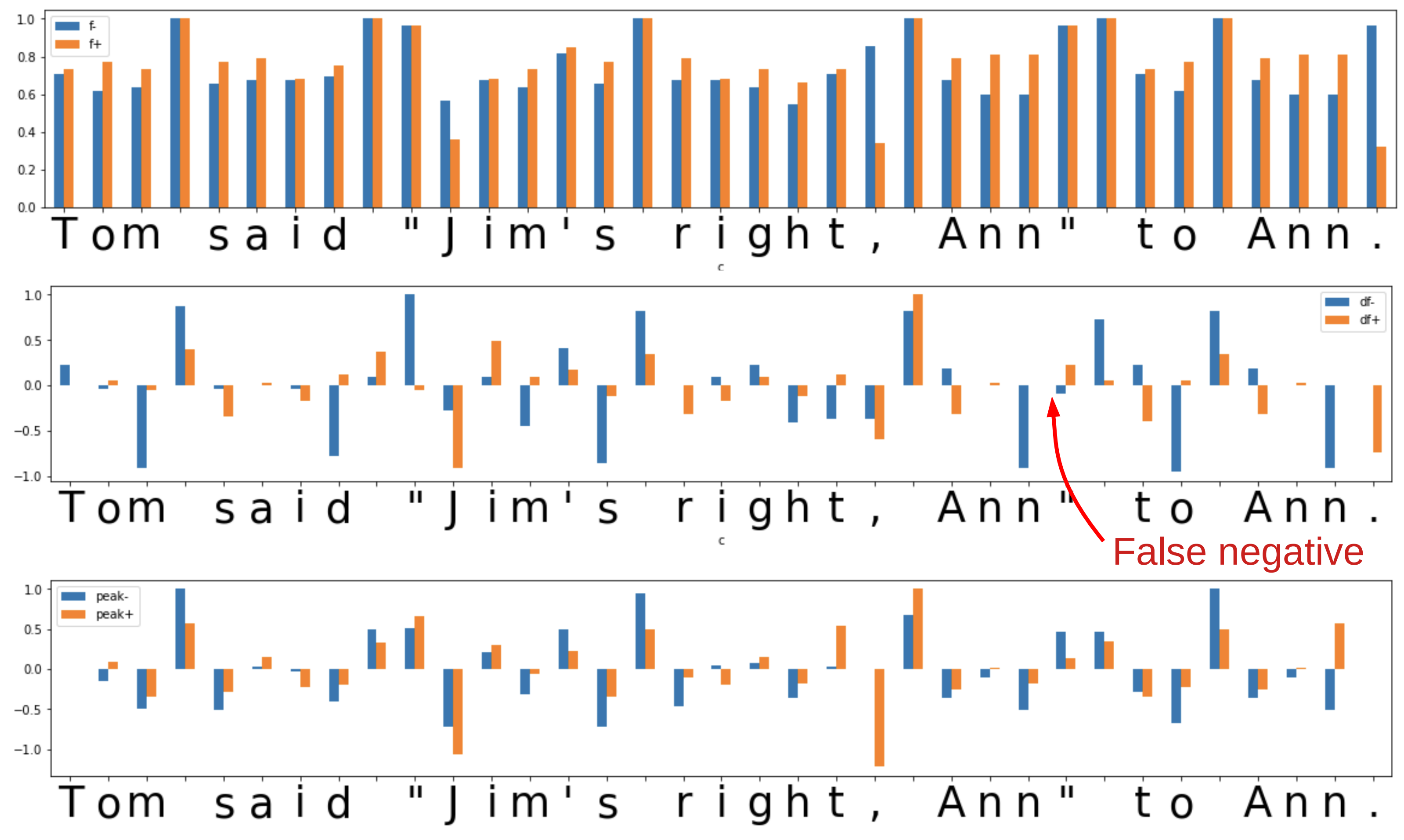

Figure 4: Using transition freedoms in forward (f+) and backward (f-) directions and their derivatives in forward (df+) and backward (df-) directions and their “peak values” (peak+ and peak-) computed on unigrams. Using plain derivatives makes punctuation marks not always separated while using “peak values” computed on derivatives identifies all words and punctuation marks correctly.

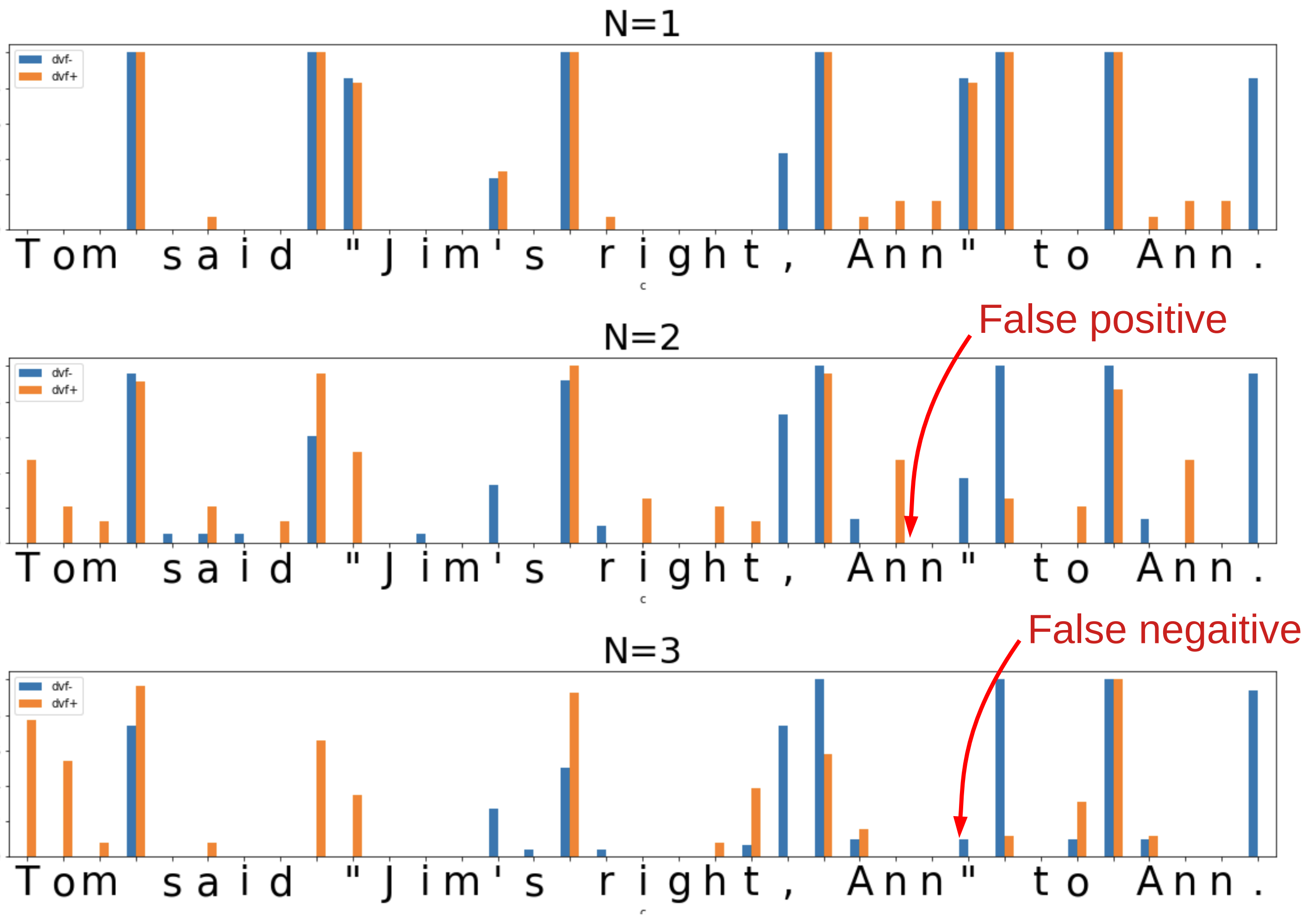

Figure 5: Deviations of transition freedoms in forward (dvf+) and backward (dvf-) directions with N-grams of varying N-values. It is clearly apparent that N=1 allows for the most accurate identification of whitespaces as well as punctuation marks. Using N > 1 does not identify all punctuation marks quite correctly and breaks some words apart.

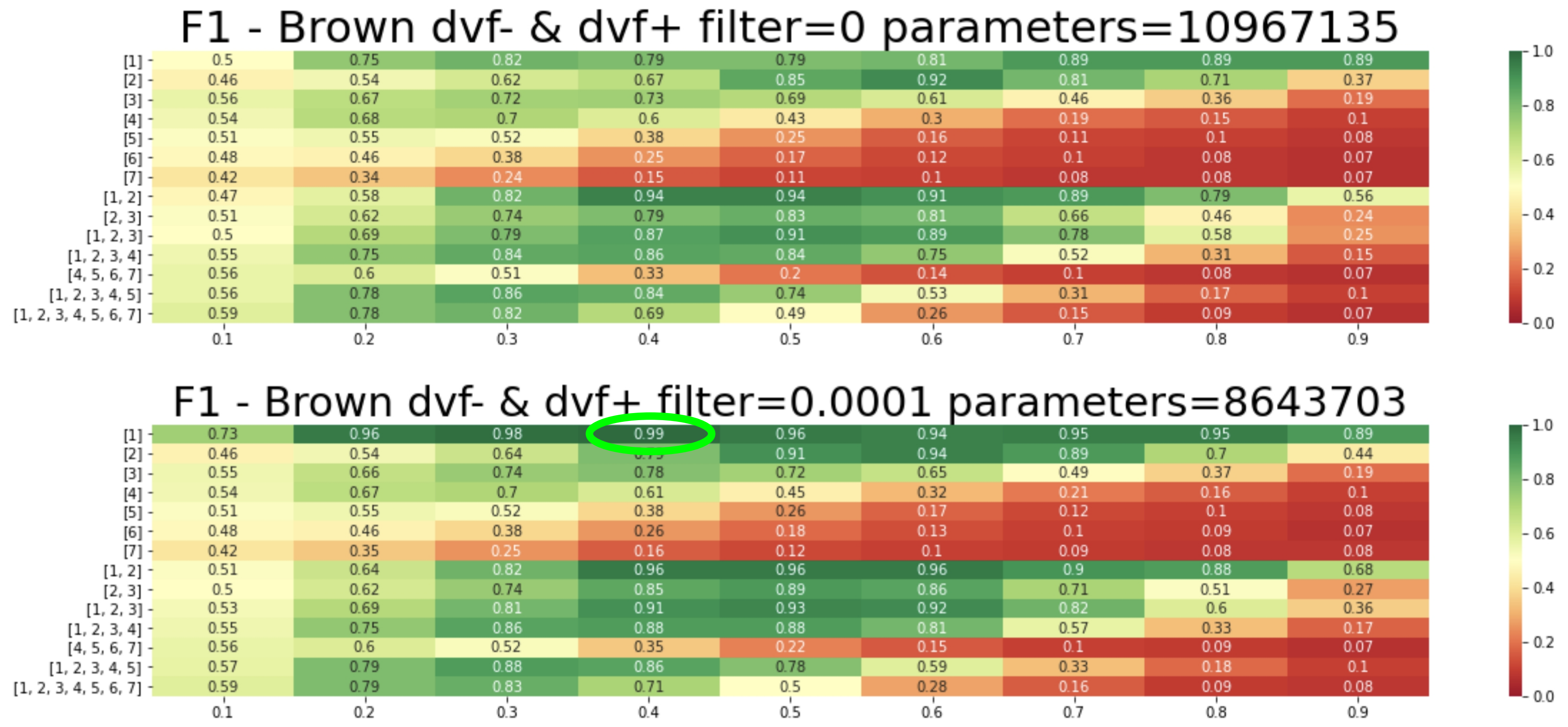

Figure 6: Heat-maps rendering F1 scores obtained for unsupervised tokenization after training on the Brown corpus with no model compression (top) and model compression with a transition frequency threshold of 0.0001 (bottom) with different combinations of N (vertical axes) and different tokenization thresholds (horizontal axes). It is seen that the highest F1 scores above 0.96 correspond to model compressed with threshold 0.0001, N = 1 (unigrams), and tokenization thresholds from 0.3 to 0.4. Model parameters are indicated in the plot titles, where each parameter corresponds to the weight or frequency count for either N-grams or transitions between N-grams.

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	-

Table: Summary of the presented research on tokenizers relying on “transition freedom” (“Freedom-based”) compared to ones based on pre-existing lexicons (“Lexicon-based”) across different languages. The middle column renders F1-score of tokenization referring to “gold-standard” rule-based tokenizers (based on hardcoded rules or hybrid tokenizers combining hardcoded rules, lexicons, and/or statistical measures such as mutual/conditional probabilities). The last column shows precision of discovery of pre-existing lexicon, with reference numbers based on rule-based tokenizers displayed in parentheses.
English: Both tokenization and lexicon discovery are solved with freedom-based tokenizers no worse than with lexicon-based ones (F1 and Precision = 0.99).
Russian: Both tokenization and lexicon discovery tasks are solved better (F1 and Precision = 1.0) with freedom-based tokenizers than with lexicon-based ones (F1 = 0.94).
Chinese: Tokenization is solved less accurately by freedom-based tokenizers than by lexicon-based ones (0.71 vs. 0.83). However, freedom-based tokenizers perform lexicon discovery relatively well compared to rule-based/hybrid tokenizers (0.92 vs. 0.94).
Preprint:https://arxiv.org/abs/2205.11443