Restoring Punctuation and Capitalization in Transcribed Speech

Agustín Gravano
agus@cs.columbia.edu

Martin Jansche
mjansche@google.com

Michiel Bacchiani
michiel@google.com

Problem

Raw ASR Output

to simulate the terrain of mars scientists put the rover in hawaii’s kilauea volcano the kids sit two thousand four hundred miles away at the nasa ames research center in mountain view california

⤳

Formatted Text

To simulate the terrain of Mars, scientists put the rover in Hawaii’s Kilauea volcano. The kids sit 2400 miles away at the NASA Ames Research Center in Mountain View, CA.

Formatting ASR output improves its readability; e.g., numbers names, disfluencies, punctuation, capitalization.

Our goal: Restore **punctuation and capitalization** (P+C) in English ASR output.

Previous studies show that textual + acoustic/prosodic models outperform purely text-based ones.

However, with massive amounts of written data and increasing computational power available, **how much can simple *n*-gram models be improved with a higher *n* and more training data?**

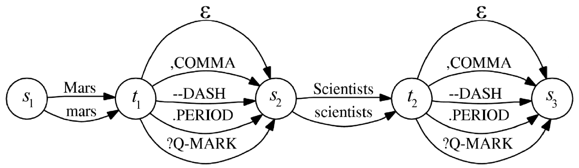
Method

Train an *n*-gram model.

Given an input string (or lattice), build its hyper-string FSA.
This FSA accepts all P+C combinations for the input string.

Compose the hyper-string FSA with the model.
This step weights each P+C combination according to the LM.

Compute the lowest-cost path.
This path corresponds to the most likely P+C combination.



Data and Experiments

Training: Internet news articles (up to 55B tokens)

Evaluation: Broadcast News (BN, 39M tokens); Wall Street Journal (WSJ, 13M tokens)

Experiments: Vary the amount of training data.
Vary the order of the *n*-gram model.
Consider two sets of punctuation symbols.

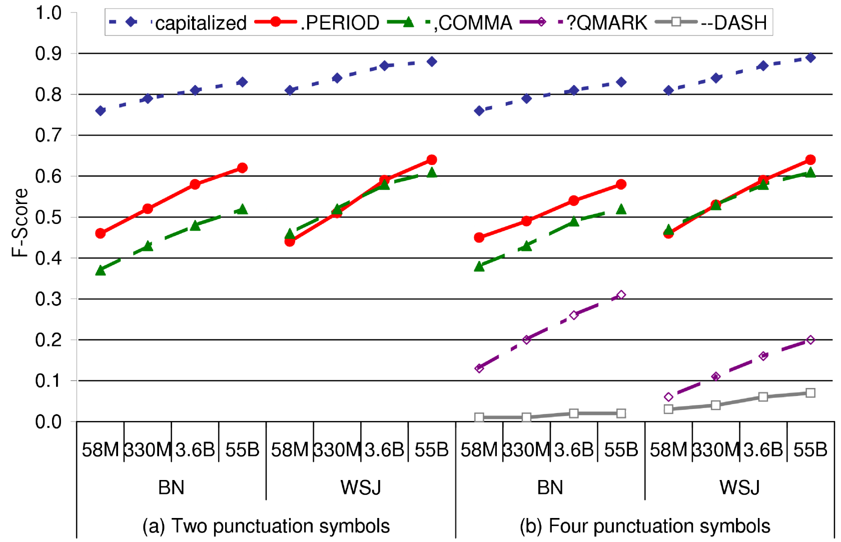
Conclusion

Capitalization is modeled well by *n*-grams; low-frequency symbols such as ?QMARK and –DASH are not.

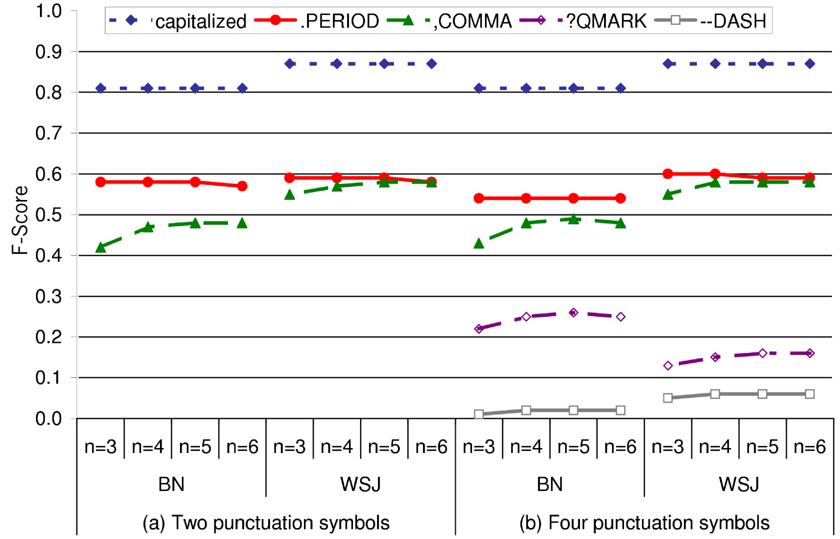
Using larger amounts of training data improves performance; increasing the order of the *n*-gram model does not.

Results

Varying the amount of training data from 58M to 55B tokens ($n = 5$)



Varying the order of the *n*-gram model from $n = 3$ to $n = 6$ (3.6B training tokens)


capitalized
.PERIOD
,COMMA
?QMARK
--DASH
F-Score
n=3
n=4
n=5
n=6
BN
WSJ
(a) Two punctuation symbols
(b) Four punctuation symbols