



BYTF: How Good Are Byte Level N-Gram F-Scores for Automatic Machine Translation Evaluation?

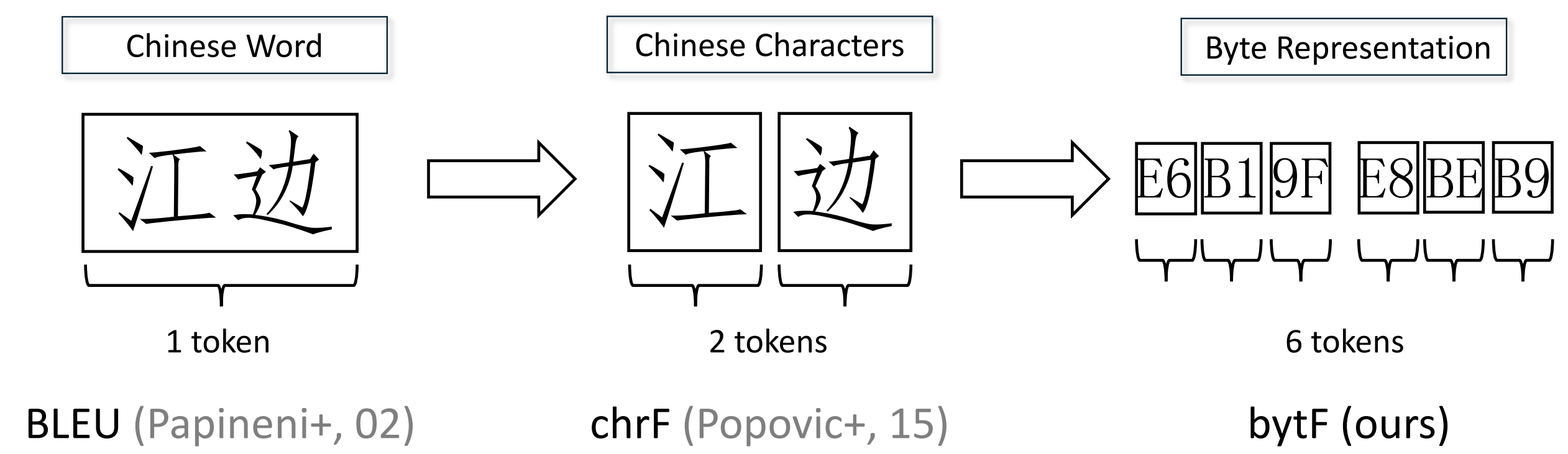
Raj Dabre, Hour Kaing, Haiyue Song

National Institute of Information and Communications Technology, Japan



Try BYTF ↑

bytF is an automatic MT quality evaluation metric that captures fine-grain information within characters



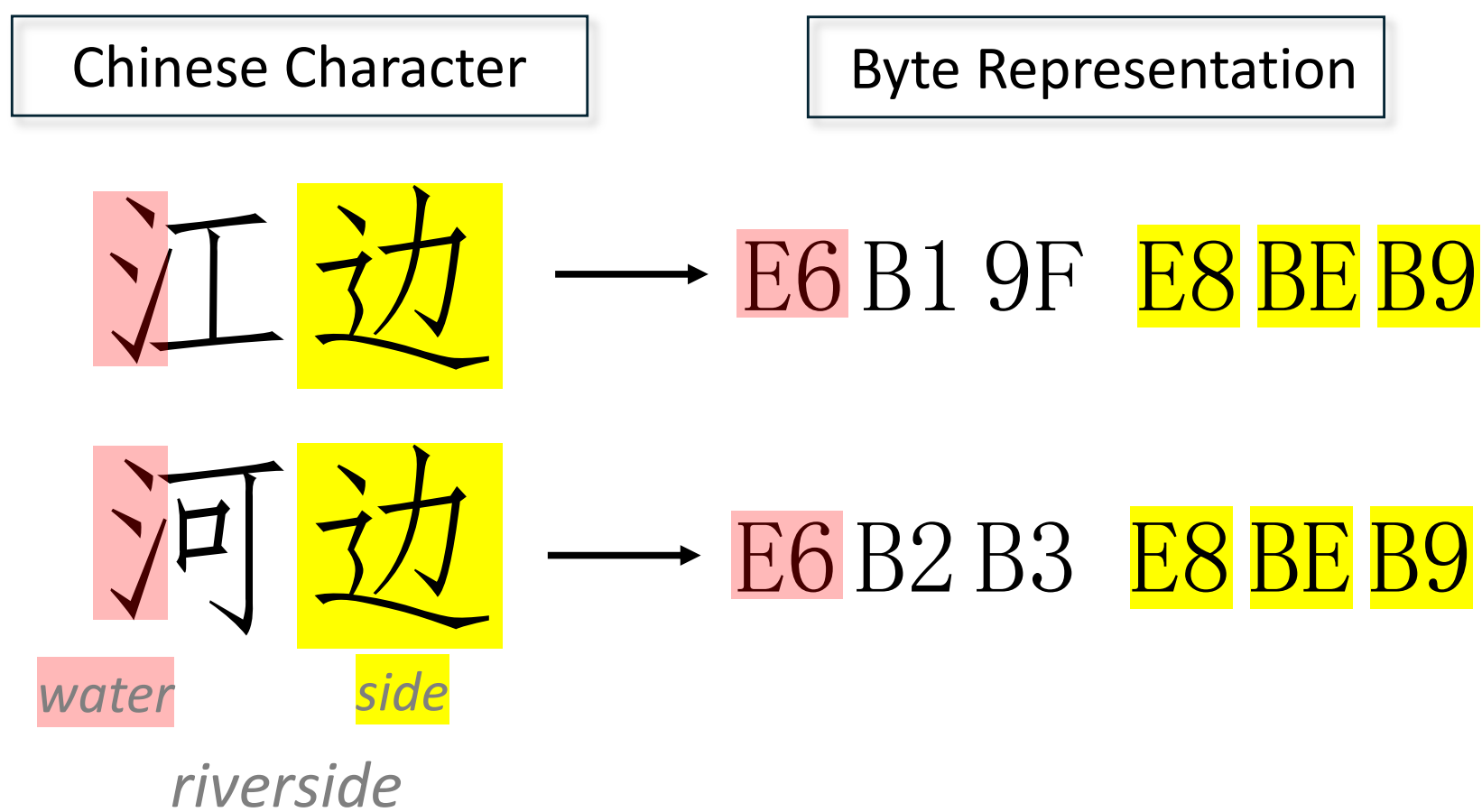
Why byte-level?

- character n-gram captures sub-word-structure similarity, chrF > BLEU
- there are sub-character-structures in Non-Latin characters
- byte n-gram captures sub-character-structure similarity, bytF > chrF > BLEU

A>B means method A shows a higher correlation with human evaluation than B

Example:

two Chinese words as **reference** and **candidate**



Precision:

word-level 1-gram: 0

char-level 1-gram: 1/2
2-gram: 0

byte-level 1-gram: 4/6
2-gram: 2/5
3-gram: 1/4
4-gram: 0

How to calculate?

Precision/Recall of a specific n -gram

- G_n is the set of byte-level n -grams

$$P_n = \frac{\sum_{g \in G_n} \min \{ \text{Count}(g, \text{cand}), \text{Count}(g, \text{ref}) \}}{\sum_{g \in G_n} \text{Count}(g, \text{cand})} \quad (3)$$
$$R_n = \frac{\sum_{g \in G_n} \min \{ \text{Count}(g, \text{cand}), \text{Count}(g, \text{ref}) \}}{\sum_{g \in G_n} \text{Count}(g, \text{ref})} \quad (4)$$

averaged Precision/Recall

$$\text{BYTP} = \frac{1}{N} \sum_{n=1}^N P_n$$

$$\text{BYTR} = \frac{1}{N} \sum_{n=1}^N R_n$$

F-score from Precision/Recall

$$\text{BYTF}_\beta = (1 + \beta^2) \frac{\text{BYTP} \cdot \text{BYTR}}{\beta^2 \text{BYTP} + \text{BYTR}}$$

$$\text{BYTF}_{+/++\beta} = (1 + \beta^2) \frac{\text{BYTP}_{+/++} \cdot \text{BYTR}_{+/++}}{\beta^2 \text{BYTP}_{+/++} + \text{BYTR}_{+/++}}$$

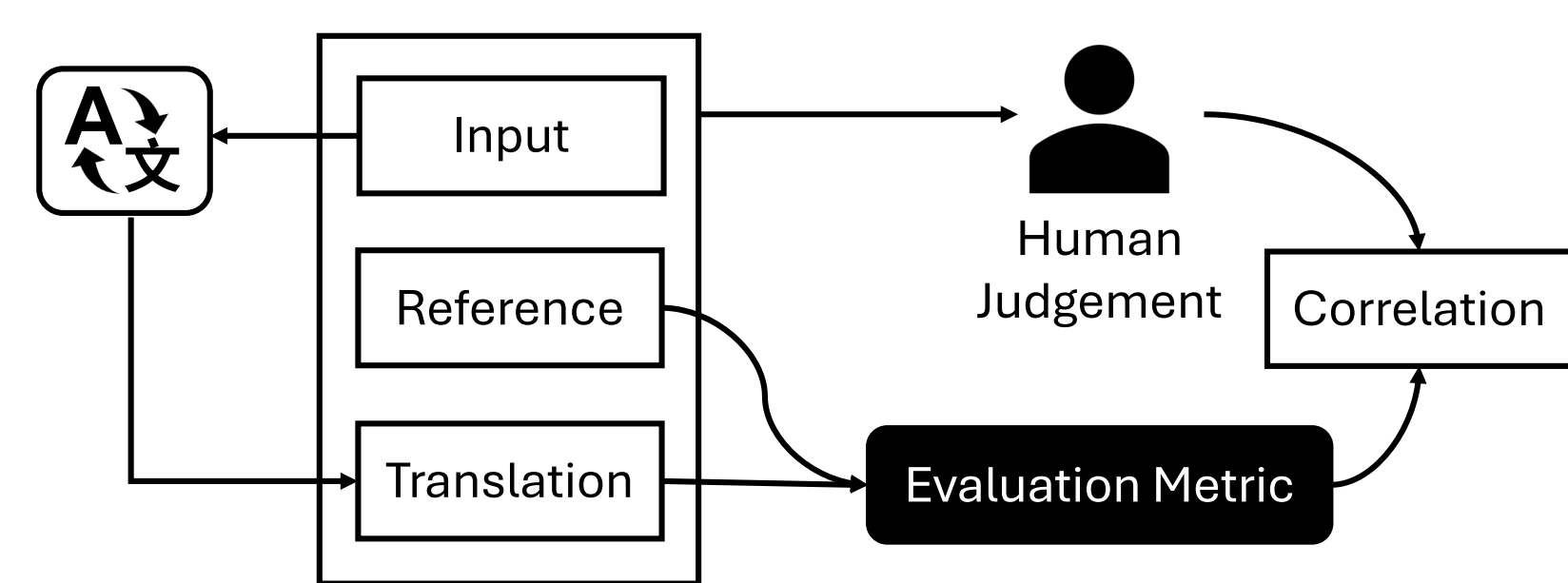
when G_n extends to byte- char- and word-level

Outperforms BLEU/chrF

Datasets:

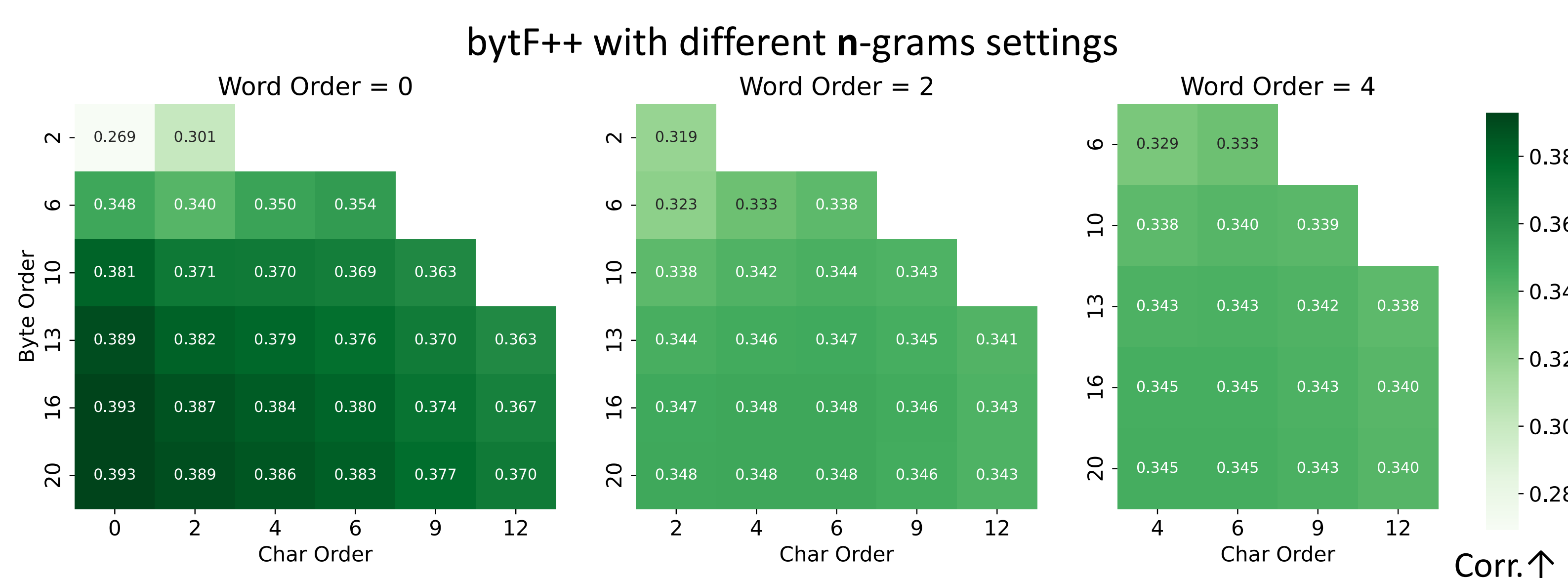
- IndicMT Eval with MQM type data
- WMT 2017-2022 with DA type data

Meta Evaluation through Correlation:



Direction	Pearson Correlation Coefficient			Kendall's Tau		
	BLEU	CHRF	BYTF	BLEU	CHRF	BYTF
Eng-Hin (IndicMT)	0.2600	0.2918 _{12,0}	0.3462 _{20,0,0†}	0.1725	0.2012 _{9,0}	0.2631 _{20,0,0†}
Eng-Guj (IndicMT)	0.2978	0.4269 _{2,0}	0.4725 _{6,0,0†}	0.2472	0.2857 _{6,0†}	0.3284 _{13,0,0†}
Eng-Mal (IndicMT)	0.2793	0.4175 _{6,0}	0.4426 _{20,0,0†}	0.3076	0.3463 _{6,2†}	0.3746 _{20,0,0†}
Eng-Tam (IndicMT)	0.2647	0.3668 _{6,0}	0.4043 _{20,0,0†}	0.2069	0.2579 _{6,0}	0.2896 _{20,0,0†}
Eng-Mar (IndicMT)	0.1954	0.2656 _{4,2†}	0.3327 _{13,0,0†}	0.1468	0.1709 _{4,2†}	0.2268 _{13,0,0†}
Ben-Hin (Wiki21)	0.0901	0.1156 _{2,0}	0.1165 _{6,4,0†}	0.0563	0.0669 _{6,0}	0.0673 _{16,9,0†}
Hin-Ben (Wiki21)	0.1116	0.1915 _{2,2}	0.1974 _{2,2,0†}	0.0956	0.1144 _{6,4}	0.1162 _{16,0,0†}
Eng-Guj (News19)	0.3992	0.4760 _{6,2†}	0.4774 _{16,6,2†}	0.2845	0.3366 _{4,0}	0.3377 _{6,6,2†}
Eng-Jpn (News20)	0.3615	0.4144 _{2,2†}	0.4213 _{2,2,2†}	0.2509	0.2769 _{2,2†}	0.2576 _{6,2,0†}
Eng-Jpn (News21)	0.2645	0.3157 _{2,2†}	0.3189 _{2,2,2†}	0.1740	0.1953 _{2,2†}	0.1895 _{2,2,2†}
Eng-Zho (News17)	0.4197	0.4717 _{2,2†}	0.4708 _{6,2,2†}	0.2951	0.3203 _{2,2†}	0.3196 _{6,2,2†}
Eng-Zho (News18)	0.3101	0.3492 _{2,2†}	0.3545 _{2,2,2†}	0.2209	0.2424 _{2,2†}	0.2444 _{2,2,2†}
Eng-Zho (News19)	0.2262	0.2481 _{2,2†}	0.2503 _{2,2,2†}	0.1350	0.1491 _{2,2†}	0.1489 _{6,2,2†}
Eng-Zho (News20)	0.2672	0.3097 _{2,2†}	0.3147 _{2,2,2†}	0.1720	0.1954 _{2,2†}	0.1962 _{2,2,2†}
Eng-Zho (News21)	0.1703	0.1834 _{2,2†}	0.1820 _{6,2,2†}	0.1050	0.1149 _{2,2†}	0.1137 _{6,2,2†}

The choice of N (in N -gram)



Variances of bytF++ and chrF++ with different settings

