Fundamenal Problems of BLEU

- BLEU overly sensitive to word forms and sequences of tokens.

Confirmed by Ref	Contains Error Flags	1-grams	2-grams	3-grams	4-grams
Yes	Yes	6.34%	1.58%	0.55%	0.29%
Yes	No	36.93%	13.68%	5.87%	2.69%
No	Yes	22.33%	41.83%	54.64%	63.88%
No	No	**34.40%**	**42.91%**	**38.94%**	**33.14%**
Total n-grams		35 531	33 891	32 251	30 611

30–40% of tokens not confirmed by reference but without errors.
⇒ Enough space for MT systems to differ unnoticed.
⇒ Low BLEU scores correlate even less.