Aryaz Eghbali
aryaz.eghbali@iste.uni-stuttgart.de

CrystalBLEU: Precisely Measuring Code Similarity

Michael Pradel
michael@binaervarianz.de

University of Stuttgart
Germany

Problem

Evaluation of Code Generating Models

1) Behavioral → Requires execution and definition of correct behavior
2) Syntactical
   a) Exact match → Downplays almost-correct results
   b) AST-based or DFG-based → Requires parsable code
   c) **BLEU score** → Accurate (enough) in practice

BLEU (Bilingual Evaluation Understudy)

- Developed for evaluating natural language translations by Papineni et al.[1] in 2002.
- Measures the precision of n-grams in the hypothesis compared to some references.

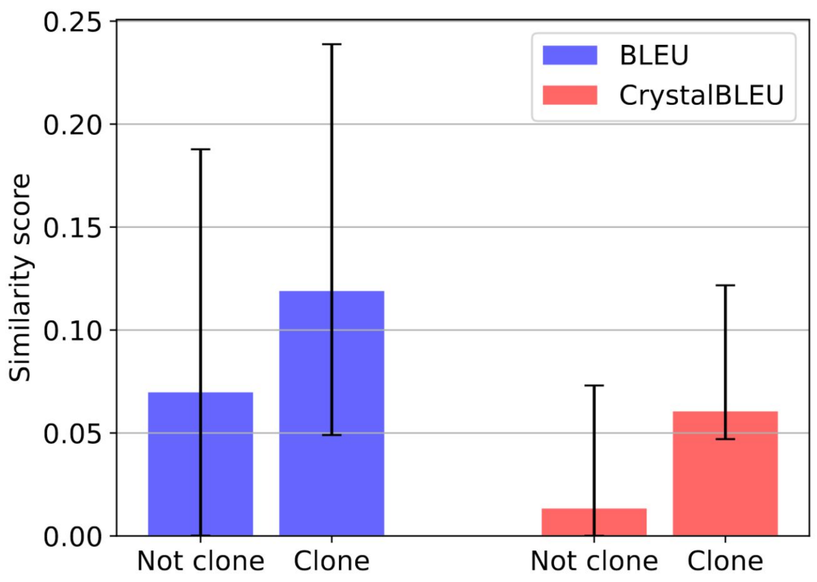
$$BLEU = exp\left(min(1 - \frac{r}{c}, 0) + \sum_{i=1}^{maxN} w_i \log p_i\right)$$

Distinguishability

Given a metric m, evaluate how well can m distinguish pairs of similar code from pairs of dissimilar code.

$Pairs_{intra}$: pairs of similar code pieces
$Pairs_{inter}$: pairs of dissimilar code pieces

$$d = \frac{m(Pairs_{intra})}{m(Pairs_{inter})}$$

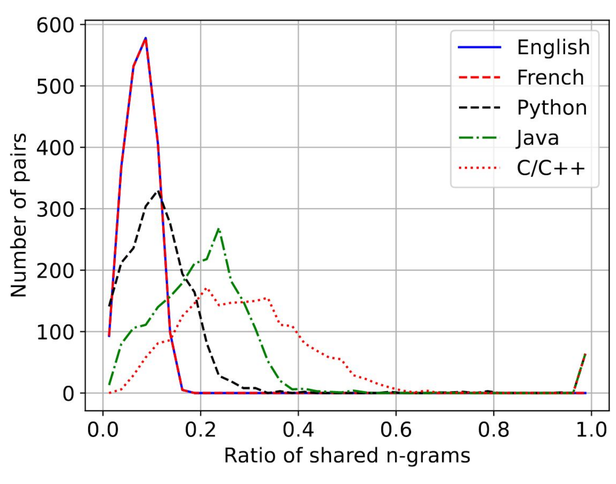


Programming and natural languages are inherently different

Top common n-grams are more frequent in NL than PL

	2-grams	% of 2-grams	4-grams	% of 4-grams
Java	) ;	5.49	) ; } }	1.34
	( )	4.75	( ) { return	1.29
	) {	3.83	( ) ; }	1.14
	; }	3.81	)) ; }	1.12
	; import	2.75	) ; } public	1.00
	))	1.73	( )) ;	0.94
	} public	1.27	) { this .	0.68
	{ return	1.24	; } public void	0.61
	} }	1.07	; } @Override public	0.54
	) .	0.99	) { if (	0.54
English	of the	2.31	” , he said	0.56
	, and	1.45	, he said ,	0.32
	in the	1.31	, of course ,	0.31
	” .	1.01	” , I said	0.29
	, the	0.85	” , she said	0.29
	to the	0.85	, he said .	0.27
	. “	0.64	” . “ I	0.22
	” ,	0.63	he said , “	0.21
	on the	0.57	” ? asked .	0.19
	, but	0.53	, I said .	0.19

Random pairs of NL share fewer n-grams than PL



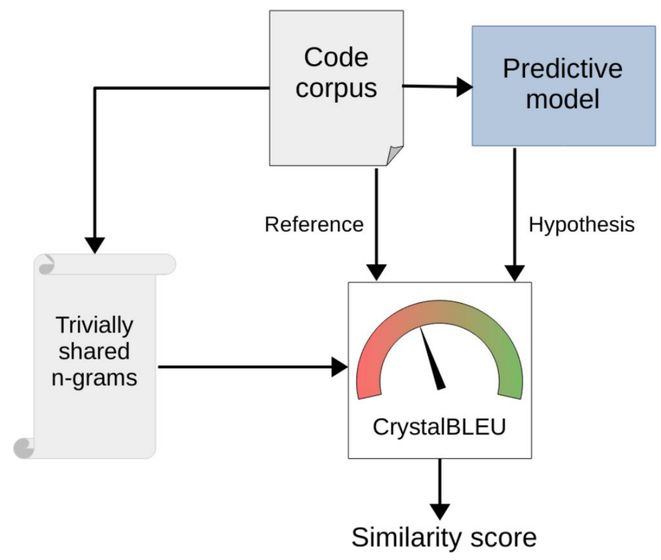
Solution

CrystalBLEU

Reduce the effects of **trivially shared n-grams**:

Preprocess (once for a dataset):
Extract common n-grams → trivially shared n-grams

1) *Remove the common n-grams from matched n-grams*
2) *Calculate BLEU with the updated matching n-grams*

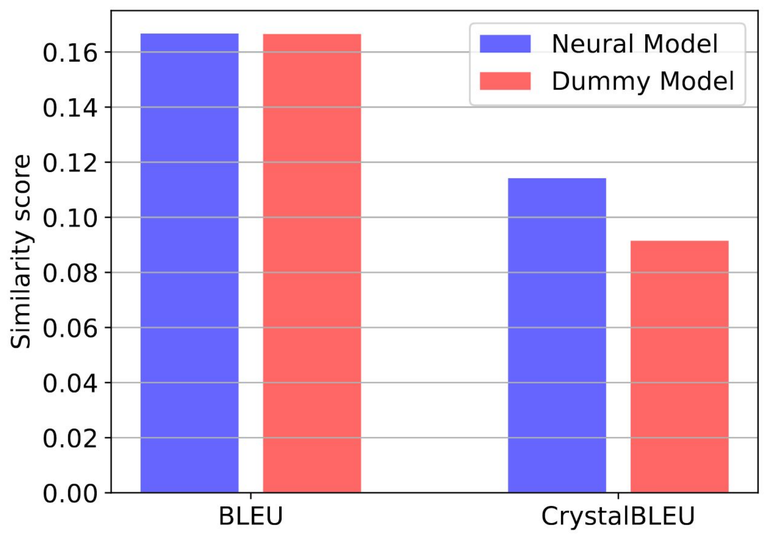


Results

Higher distinguishability

Language	BLEU	CodeBLEU	**CrystalBLEU**
ShareCode Java	2.47	1.44	**6.50**
ShareCode C++	2.82	N/A	**8.29**
BigCloneBench	1.44	1.18	**2.77**

CrystalBLEU correctly scores models that generate common n-grams lower than better models. In contrast, BLEU can misrepresent.



CrystalBLEU Features

Property	BLEU	CodeBLEU	RUBY	CrystalBLEU
Language-agnostic	✅	❌	❌	✅
Handle incomplete and partially incorrect code	✅	❌	❌	✅
Efficient	✅	❌	❌	✅
Correlate well with human judgment	✅	✅	✅	✅
High distinguishability	❌	❌	N/A	✅

References

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Annual Meeting on Association for Computational Linguistics (ACL). 311–318.