Performance Metrics for Different Tokenization Schemes and Model Architectures

Model	**Prec.**	**Rec.**	**F1**	**BLEU**
Baseline	0.8957	0.8923	0.8934	73.69
+ BPE Tokenizer	0.9571	0.9475	0.9517	90.29
+ Unigram Tokenizer	0.9614	0.9508	0.9553	90.48
+ WordPiece Tokenizer (WP)	**0.9667**	0.9513	0.9579	92.30
+ WP + Linked Attention	0.9660	0.9503	0.9570	92.21
CASTLE: WP + Linked + KG	0.9718	**0.9559**	**0.9629**	**92.72**
BART-large (finetuned)	0.9641	0.9558	0.9592	90.83

Performance of GEC Models by Error Category (F1 and BLEU Scores)

Model	Semantics			Morphology			Syntax		
	Ambig.	Diction	Pleon.	Affix.	Word F.	Redup.	Complt.	Prep.	Phrase S.
F1-Score									
Baseline	0.8941	0.8933	0.8906	0.8941	0.8923	0.8935	0.8928	0.8935	0.8939
+ BPE	0.9519	0.9510	0.9509	0.9518	0.9514	0.9523	0.9517	0.9517	0.9520
+ Unigram	0.9550	0.9541	0.9549	0.9559	0.9551	0.9551	0.9555	0.9546	0.9559
+ WordPiece	0.9582	0.9575	0.9567	0.9578	0.9576	0.9591	0.9578	0.9582	0.9579
+ WP + Linked	0.9502	0.9500	0.9490	0.9506	0.9507	0.9518	0.9506	0.9510	0.9509
CASTLE	**0.9669**	0.9359	**0.9652**	**0.9611**	**0.9623**	**0.9685**	0.9355	**0.9659**	**0.9632**
BART-large	0.9594	**0.9585**	0.9579	0.9593	0.9591	0.9592	**0.9594**	0.9595	0.9593
BLEU-Score									
Baseline	73.84	73.75	73.30	73.81	73.37	73.41	73.54	73.58	73.97
+ BPE	90.26	90.00	90.40	90.36	90.27	90.12	90.21	90.23	90.52
+ Unigram	90.24	90.01	90.63	90.59	90.36	90.42	90.60	90.30	90.81
+ WordPiece	**92.19**	**91.80**	91.91	92.51	92.22	92.30	**92.14**	**92.12**	**92.67**
+ WP + Linked	91.87	91.50	91.65	92.19	91.96	92.03	91.85	91.85	92.41
CASTLE	92.05	90.67	**96.51**	**93.93**	**96.37**	**96.22**	88.52	91.66	92.12
BART-large	90.75	90.35	90.32	91.08	90.67	90.55	90.76	90.80	91.15