Evaluation of HTR models without Ground Truth Material

Phillip Benjamin Ströbel, Simon Clematide, Martin Volk, Raphael Schwitter, Tobias Hodel, David Schoch (2022)

Victoria Beltrán Domínguez

Índice

1. Introducción
2. Trabajos relacionados
3. Métricas
4. Experimentos
5. Resultados y discusión
6. Conclusiones

1. Introducción

- Mejora en modelos HTR gracias a las redes neuronales
- No existe un uso más amplio en bibliotecas y archivos
- El despliegue de modelos HTR viene con incertidumbres
- Alternativas:
  - Recopilar nuevo GT
  - Puntajes de confianza
  - Encontramos métricas de evaluación sin GT

2. Trabajos relacionados

- Muchos apuntan a la dificultad de la tarea

- Pocos estudios examinan la evaluación en producción de HTR:
  - Intervalos de confianza como medida de predicción:
    - Sarkar et al. (2001) utiliza intervalos de confianza para construir un sistema de triaje de documentos.
    - Springmann et al. (2016) demuestra la relación entre precisión e intervalos de confianza.
  - Alex y Burns demuestran la importancia del lexicón.

3. Métricas

Encontramos dos tipos de métricas:

a. Basadas en el lexicón

b. Basadas en la perplejidad

3. Métricas: basadas en lexicón

Origen:

- Alex y Burns (2014) indicaron que una proporción simple de palabras reconocidas contra un léxico se correlaciona bien con el juicio humano sobre la calidad de OCR.

- Clausner et al. (2016) mostró que las características léxicas son las más adecuadas para predecir la calidad de OCR.

3. Métricas: basadas en lexicón

1. **Token ratio**: porcentaje de tokens reconocidos por el modelo HTR que también ocurren en una referencia .

$$\mathcal{R}_{\text{token}} = \frac{c(T \in \mathcal{V})}{c(T)}.$$

2. **Character N-gram ratio**: porcentaje de n-gramas reconocidos por el modelo HTR que también ocurren en una referencia.

$$\mathcal{R}_{n-\text{gram}} = \frac{c(N \in \mathcal{G})}{c(N)}.$$

3. Métricas: basadas en perplejidad

Origen:

- Los LM (modelos de lenguaje) forman parte integral del proceso de HTR
- Se sugiere utilizar LM externos para evaluar los resultados del modelo de HTR.
- Primer enfoque que utiliza LM para la estimación de la calidad de HTR

3. Métricas: basadas en perplejidad

1. **Perplejidad estadística del LM (PPL):**

- Tarea de un modelo de lenguaje: predecir palabra máxima verosimilitud para n-gramas:

$$P_{MLE}(w_n|w_1...w_{n-1}) = \frac{c(w_1...w_n)}{c(w_1...w_{n-1})}$$

- Mide la sorpresa del modelo al calcular la probabilidad de la secuencia de palabras.

$$PPL(W) = 2^{H(W)} \qquad H(W) = -\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i)$$

3. Métricas: basadas en perplejidad

2. Pseudo perplejidad (PPPL):

- No se puede aplicar directamente PPL a transformers como BERT o Roberta por estar enmascarados. Utilizamos PPPL:

$$\mathrm{PPPL}(\mathbb{W}) := \exp\left(-\frac{1}{N}\sum_{\boldsymbol{W}\in\mathbb{W}}\mathrm{PLL}(\boldsymbol{W})\right)$$

$$\mathrm{PLL}(\boldsymbol{W}) := \sum_{t=1}^{|\boldsymbol{W}|}\log P_{\mathrm{MLM}}(\boldsymbol{w}_t \mid \boldsymbol{W}_{\setminus t};\Theta).$$

4. Experimentos

Idea:

1. Utilizar datos que tienen GT
2. Entrenar diferentes modelos
3. Calcular las métricas explicadas para cada modelo
4. Comprobar si existe correlación entre el CER y las métricas

4. Experimentos

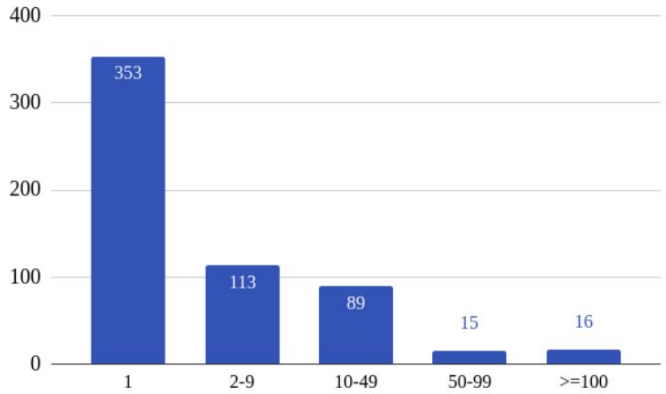
Idea:

1. Utilizar datos que tienen GT
2. Entrenar diferentes modelos
3. Calcular las métricas explicadas para cada modelo
4. Comprobar si existe correlación entre el CER y las métricas

4. Experimentos: datos

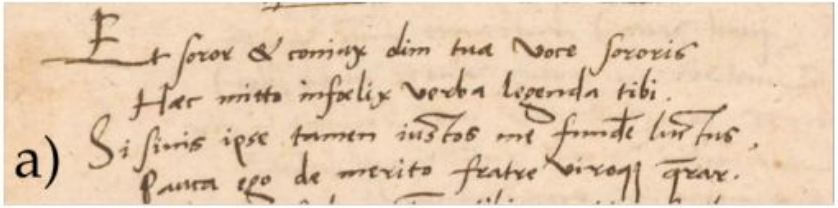
Principalmente dos conjuntos de datos:

1. Correspondencia de Heinrich Bullinger:
    - Latín
    - Alemán
2. Volumen escrito por Rudolf Gwalter

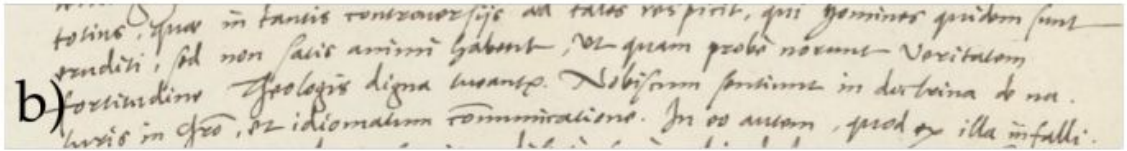


4. Experimentos: datos

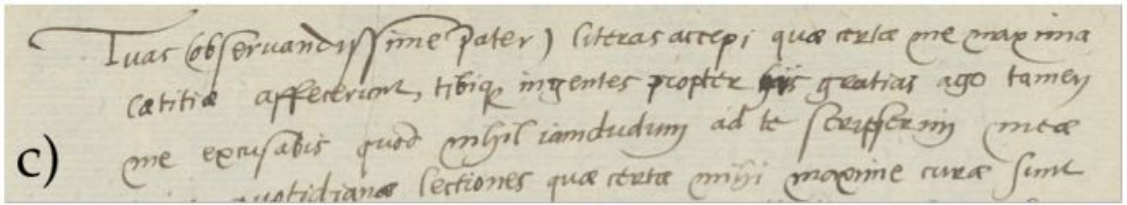
a)



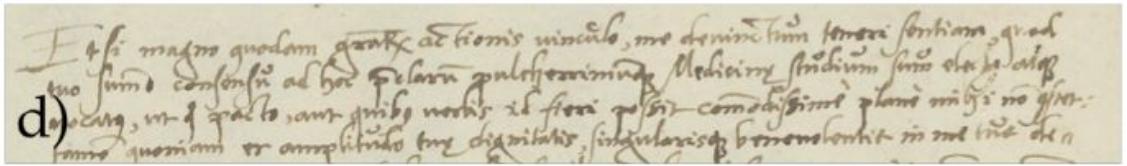
b)



c)



d)



4. Experimentos: datos

Particiones de train:

1. Correspondencia de Heinrich Bullinger (~20000 líneas - 10% validación)
2. Volumen escrito por Rudolf Gwalter (~4000 líneas - 10% validación)

Particiones de test:

1. 825 líneas de cartas de Bullinger no vistas
2. 57 líneas de cartas de Rudolf Gwalter

4. Experimentos

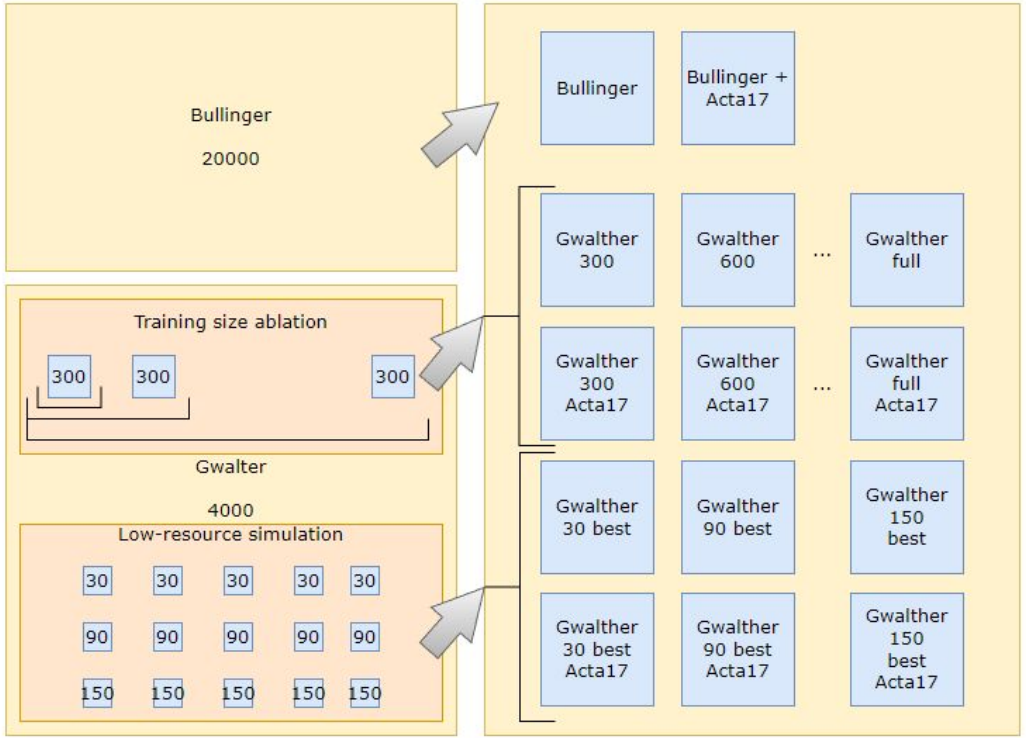
Idea:

1. Utilizar datos que tienen GT
2. Entrenar diferentes modelos
3. Calcular las métricas explicadas para cada modelo
4. Comprobar si existe correlación entre el CER y las métricas

4. Experimentos: modelos

32 modelos

(Acta17 HTR+):



4. Experimentos: modelos

	Number of lines for training														
	30	90	150	300	600	900	1200	1500	1800	2100	2400	2700	3000	3300	3600
No base	39.33	14.37	11.39	7.28	5.55	4.96	4.36	4.14	3.9	3.94	3.64	3.59	3.36	3.24	3.29
Acta-based	13.67	8.38	7.08	5.5	4.42	4.03	3.66	3.43	3.27	3.13	3.09	3.2	2.82	2.8	2.74
Other models (evaluated on the "Gwalther" validation set)															
Bullinger	6.99														
Bullinger+Acta	6.56														
Acta_17	14.66														
Spruchakten	15.95														

4. Experimentos

Idea:

1. Utilizar datos que tienen GT
2. Entrenar diferentes modelos
3. Calcular las métricas explicadas para cada modelo
4. Comprobar si existe correlación entre el CER y las métricas

4. Experimentos: métricas

Requisitos:

- **Token ratio**: tokens de referencia
- **Character N-gram ratio**: n-gramas de referencia
- **PPL**: modelo clásico
- **PPPL**: transformer

4. Experimentos: métricas

Solución:

- Corpus CC-100 (Latin):
  - ~206 millones de tokens (preprocesados)
  - Partición del 90% entrenamiento y 10% para test
  - LM estadístico: 5- gramas con interpolación Kneser-Ney
- BERT pre entrenado multilingüe
- RoBERTa

4. Experimentos

Idea:

1. Utilizar datos que tienen GT
2. Entrenar diferentes modelos
3. Calcular las métricas explicadas para cada modelo
4. Comprobar si existe correlación entre el CER y las métricas

4. Experimentos: correlación

Tres aspectos:

1. **Habilidad de estimar calidad:**

Hipótesis nula H0: las métricas y los CER no se correlacionan
Ajustar de modelos lineales

2. **Habilidad de ordenar modelos:**

Hipótesis nula H0: no existe una correlación entre las clasificaciones de diferentes modelos basadas en CER y las basadas en las métricas

Coeficiente de correlación de ranking de Spearman $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$

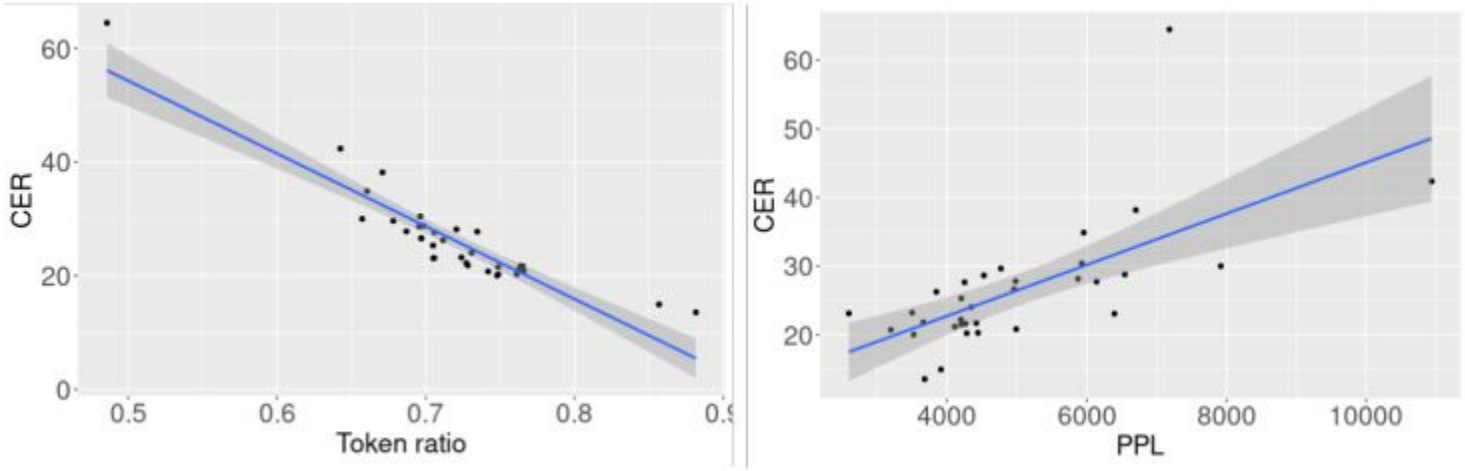
3. **Habilidad de predecir mejor modelo en top 1, top 3 y top 5**

5. Resultados y discusión

Habilidad de estimar calidad

Metrics		Adj. R^2		
		$Gwalther_{433}$	$Bullinger_{825}$	$Gwalther_{57}$
PPPL	BERT	0.99^{2}	0.80^{2}	0.92^{2}
	RoBERTa	0.98^{1}	0.69^{2}	0.88^{2}
PPL	Statistical LM	0.65^{2}	0.52^{1}	0.42^{1}
Token ratio		0.99^{1}	0.95^{4}	0.92^{3}
Character *n*-grams	2-gram	0.42^{2}	-0.01^{1}	0.14^{2}
	3-gram	-0.03^{1}	0.09^{1}	-0.03^{1}
	4-gram	0.80^{2}	0.35^{1}	0.06^{1}
	5-gram	0.96^{2}	0.79^{1}	0.81^{1}
	6-gram	0.99^{2}	0.96^{3}	0.96^{2}
	7-gram	0.99^{2}	0.99^{2}	0.97^{2}

5. Resultados y discusión



5. Resultados y discusión

Habilidad de ordenar modelos (Spearman).Significance levels: 0.05, 0.025, **0.01**, 0.005, 0.001

Metrics		Ranking Reference (CER)		
		Gwalther_{433}	Bullinger_{825}	Gwalther_{57}
PPPL	BERT	*0.98*	*0.90*	*0.90*
	RoBERTa	*0.96*	*0.82*	*0.85*
PPL	Statistical LM	*0.78*	*0.65*	*0.71*
Token ratio		*0.98*	*0.95*	*0.91*
Character *n*-grams	2-gram	-0.28	0.28	0.13
	3-gram	0.01	*0.48*	*0.36*
	4-gram	*0.58*	*0.62*	0.15
	5-gram	*0.90*	*0.88*	*0.70*
	6-gram	*0.97*	*0.94*	*0.92*
	7-gram	*0.99*	*0.97*	*0.94*

5. Resultados y discusión

Habilidad de predecir mejor modelo en top 1, top 3 y top 5

Metrics		Ranking reference		
		Gwalther_{433}	Bullinger_{825}	Gwalther_{57}
PPPL	BERT	1	1	1
	RoBERTa	1	1	1
PPL	Statistical LM		3	
Token ratio		3	1	1
Character *n*-grams	2-gram			
	3-gram		3	5
	4-gram		1	3
	5-gram	5	1	1
	6-gram	1	1	1
	7-gram	3	1	1

6. Conclusiones

- Presentación de diferentes métricas para evaluar modelos sin GT
- Propuesta de diferentes experimentos que han conseguido demostrar la relevancia de cada una de las métricas
- Elección de los mejores modelos en cada caso