Reconocimiento multimodal: Combinando escritura manuscrita y habla

Doctorado en informática

Autor:
Emilio Granell Romero
egranell@dsic.upv.es

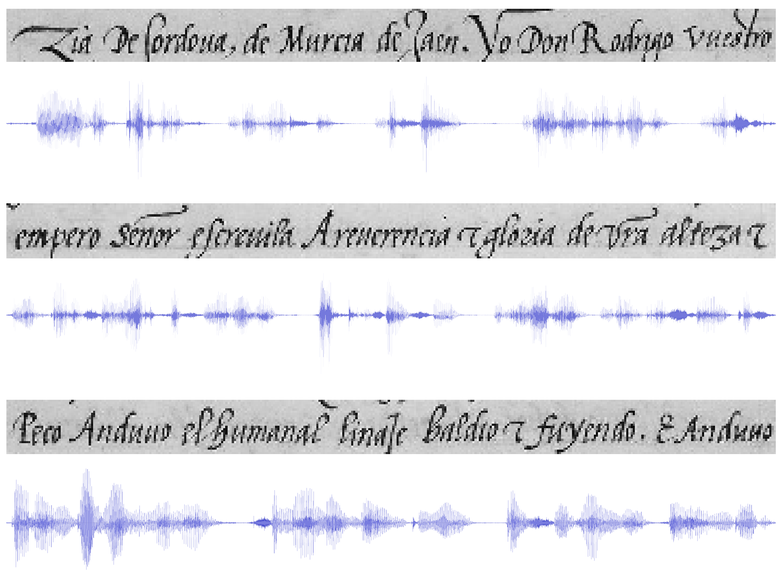
Supervisor:
Dr. Carlos D. Martínez Hinarejos
cmartine@dsic.upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Introducción

Para poder acceder a la información contenida en los documentos digitales de texto histórico es necesario disponer de su transcripción. Estas transcripciones se pueden obtener mediante el uso del reconocimiento del texto manuscrito en las páginas digitalizadas o utilizando el reconocimiento automático del habla en el dictado de los contenidos.

En este trabajo, estamos comprobando la eficacia de una tercera opción, que es utilizar ambos sistemas en una combinación multimodal.



Posibles aplicaciones

Mejorar los sistemas de transcripción asistida por ordenador de textos manuscritos.

Mejorar el reconocimiento en ambientes ruidosos al combinar las salidas de varios sistemas de una misma modalidad. (Por ejemplo, el reconocimiento del habla en vehículos).

Referencias

- S. Ishimaru, H. Nishizaki and Y. Sekiguchi,"Effect of Confusion Network Combination on Speech Recognition System for Editing", *Proc. 3rd APSIPA ASC*, pp. 1–4, 2011.
- V. Romero, L. A. Leiva, A. H. Toselli and E. Vidal, "Interactive multimodal transcription of text images using a web-based demo system", *In Proc. 14th ACM IUI* pp. 477-478, 2009.

Objetivos general y específicos

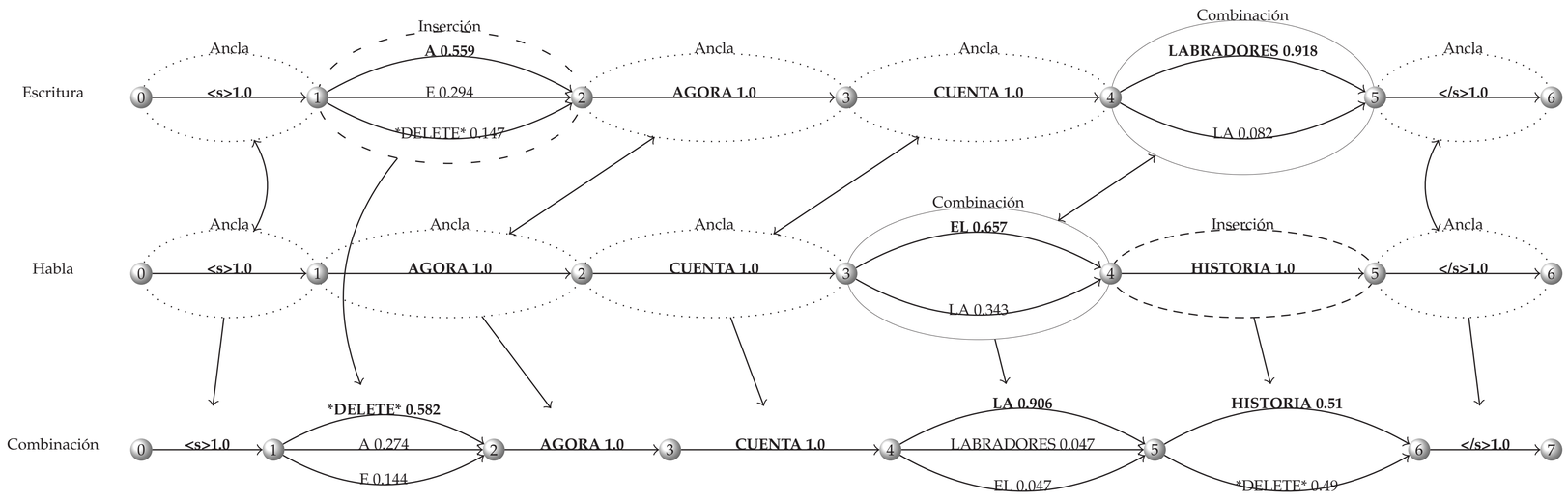
Objetivo general

Reducir el error en el reconocimiento de textos manuscritos.

Objetivos específicos

1. Estudiar las principales técnicas de modelado óptico y acústico,
2. la adaptación de los modelos al escritor y/o locutor,
3. las técnicas de interacción multimodal iterativa,
4. las técnicas de combinación de las hipótesis de salida, y
5. diseñar una nueva técnica de combinación multimodal.

Combinando las salidas de múltiples reconocedores



- Idea: Combinar salidas de distintas modalidades para corregir errores.
- Formalismo: Por el teorema de Bayes y asumiendo independencia entre salidas:

$$\Pr(Escritura, Habla|palabra) \simeq \Pr(Escritura|palabra)\Pr(Habla|palabra) \quad (1)$$

- Implementación: Versión ponderada por parámetro α:

$$\Pr(Escritura, Habla|palabra) \simeq \Pr(Escritura|palabra)^{\alpha}\Pr(Habla|palabra)^{1-\alpha} \quad (2)$$

α permite equilibrar la confianza relativa de cada sistema.

Experimentos realizados y resultados obtenidos

Utilizando dos páginas del corpus de escritura manuscrita RODRIGO y la colaboración de siete locutores, hemos realizado tres experimentos con esta nueva técnica de combinación[a]:

1. Combinación multimodal Iterativa[b].

Referencia		Resultados	
WER	CER	WER	CER
45.1%	23.6%	38.0%	17.4%

2. Combinación unimodal y multimodal.

Referencia		Res. unimodal		Res. multimodal	
WER	CER	WER	CER	WER	CER
32.9%	15.7%	31.1%	13.3%	28.2%	13.1%

3. Aplicación en un entorno de transcripción interactiva de texto manuscrito (CATTI)[c].

Referencia		Resultados	
WSR	WCR	WSR	WCR
29.0%	0.88	22.5%	0.78

[a] Trabajos presentados a los congresos internacionales: ICDAR 2015, CAIP 2015 y EMNLP 2015, respectivamente.

[b] WER (Word Error Rate) y CER (Character Error Rate): representan el error de una transcripción a nivel de palabra y de carácter, correspondientemente.

[c] WSR (Word Stroke Ratio) y WCR (Word Click Ratio): representan el nivel de error a la salida del sistema CATTI y el número de acciones de ratón adicionales que el usuario debe realizar para alcanzar dicho valor de error.

Conclusiones y trabajo futuro

- Los experimentos realizados confirman que la combinación de las salidas de diferentes reconocedores permite mejorar y acelerar la transcripción automática de textos históricos manuscritos, reduciendo de una forma considerable el esfuerzo humano necesario para obtener la transcripción completa.
- Como trabajo futuro, nos proponemos refinar esta técnica de combinación, así como experimentar con ella con nuevos conjuntos de datos y diferentes aplicaciones. (Por ejemplo, combinar el reconocimiento de texto manuscrito on-line y off-line).