Tecnologías para al análisis y metadatado de contenidos audiovisuales

Eduardo Lleida
Instituto de Investigación en Ingeniería de Aragón
Universidad de Zaragoza
lleida@unizar.es

Cátedra RTVE - UNIZAR

Fecha de creación: 10 de Julio de 2017

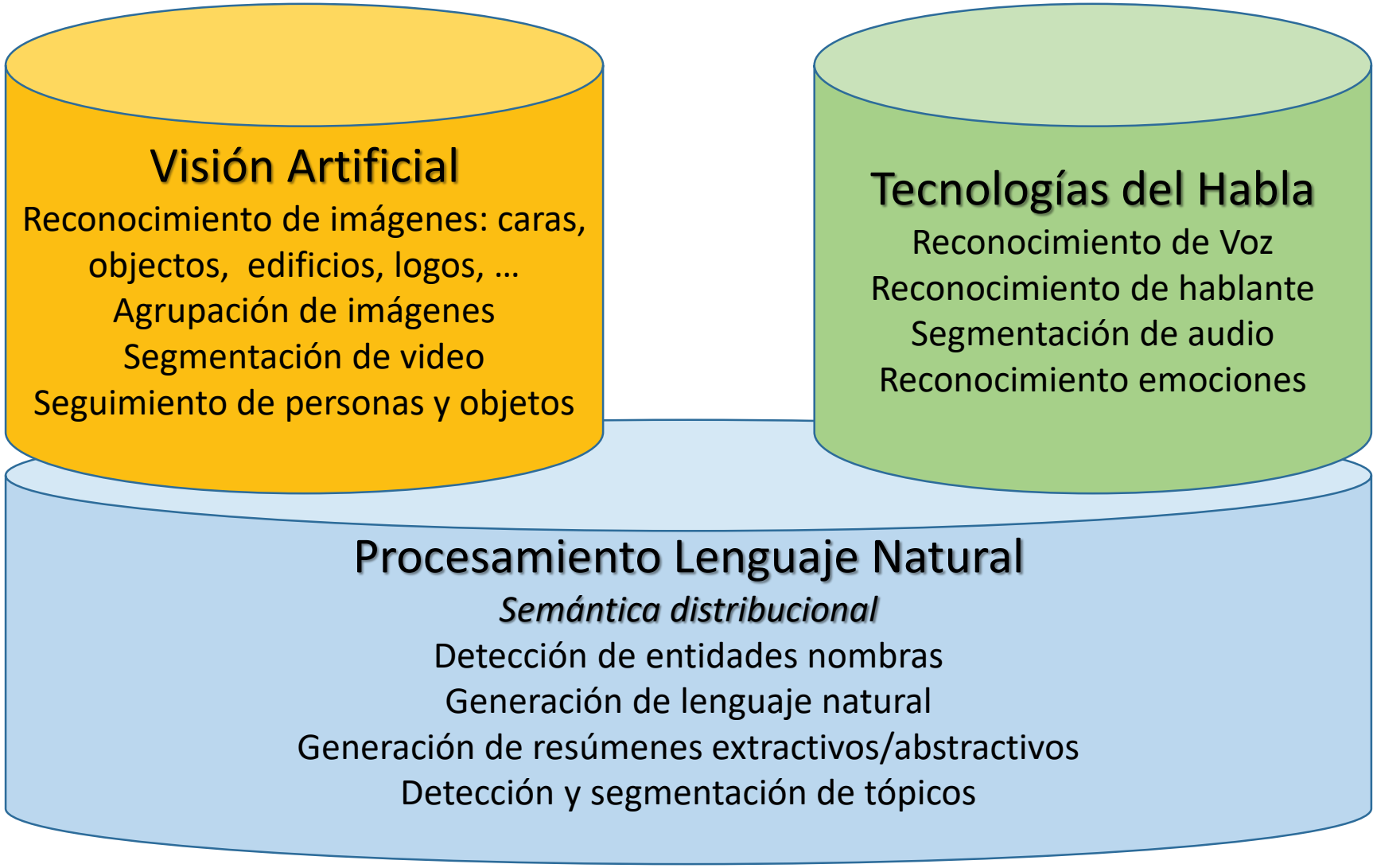
Objetivo:

realización de actividades de formación, investigación, estudio y divulgación en el área de las Tecnologías de la Información y de las Comunicaciones relacionadas con el Big Data aplicado al análisis de los contenidos audiovisuales y sonoros.

Actividades:

- herramientas para
  - el análisis del contenido partiendo de transcripción a texto de las grabaciones
  - reconocimiento de voces, caras, edificios emblemáticos y logotipos y su ubicación en el time line de la media
  - descripción automática de imágenes, planos y secuencias, en lenguaje natural
  - creación de resúmenes de forma automática
- Lanzamiento de un reto anual

Tecnologías para el análisis y metadatado de contenidos audiovisuales



Tecnologías para el análisis y metadatado de contenidos audiovisuales

Análisis y descripción del contenido audiovisual:

¿Qué dice?	Reconocimiento del habla
¿Quién habla?	Reconocimiento del hablante
¿Quién y cuándo habla?	Diarización
¿Cómo lo dice?	Reconocimiento emociones
¿Quién aparece en la imagen?	Reconocimiento facial
¿Qué aparece en la imagen?	Reconocimiento del entorno físico
	Reconocimiento de objetos
	Reconocimiento de logos, OCR, …
¿Qué hay en el audio?	Segmentación audio (voz,música,…)

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Análisis y descripción del contenido audiovisual:

Descripción del contenido audiovisual

- Detección y segmentación de tópicos
- Detección de entidades nombradas
- Descripción de imagen/vídeo
- Generación automática de resúmenes textuales
  - Extractivos vs Abstractivos
- Generación automática resúmenes audiovisuales
  - Estático (fotogramas) vs Dinámico (secuencia de tomas)

Espacios semánticos

Semántica Distribucional

*similitud **semántica** entre unidades léxicas puede detectarse a través de la búsqueda de coincidencias en el contexto lingüístico*

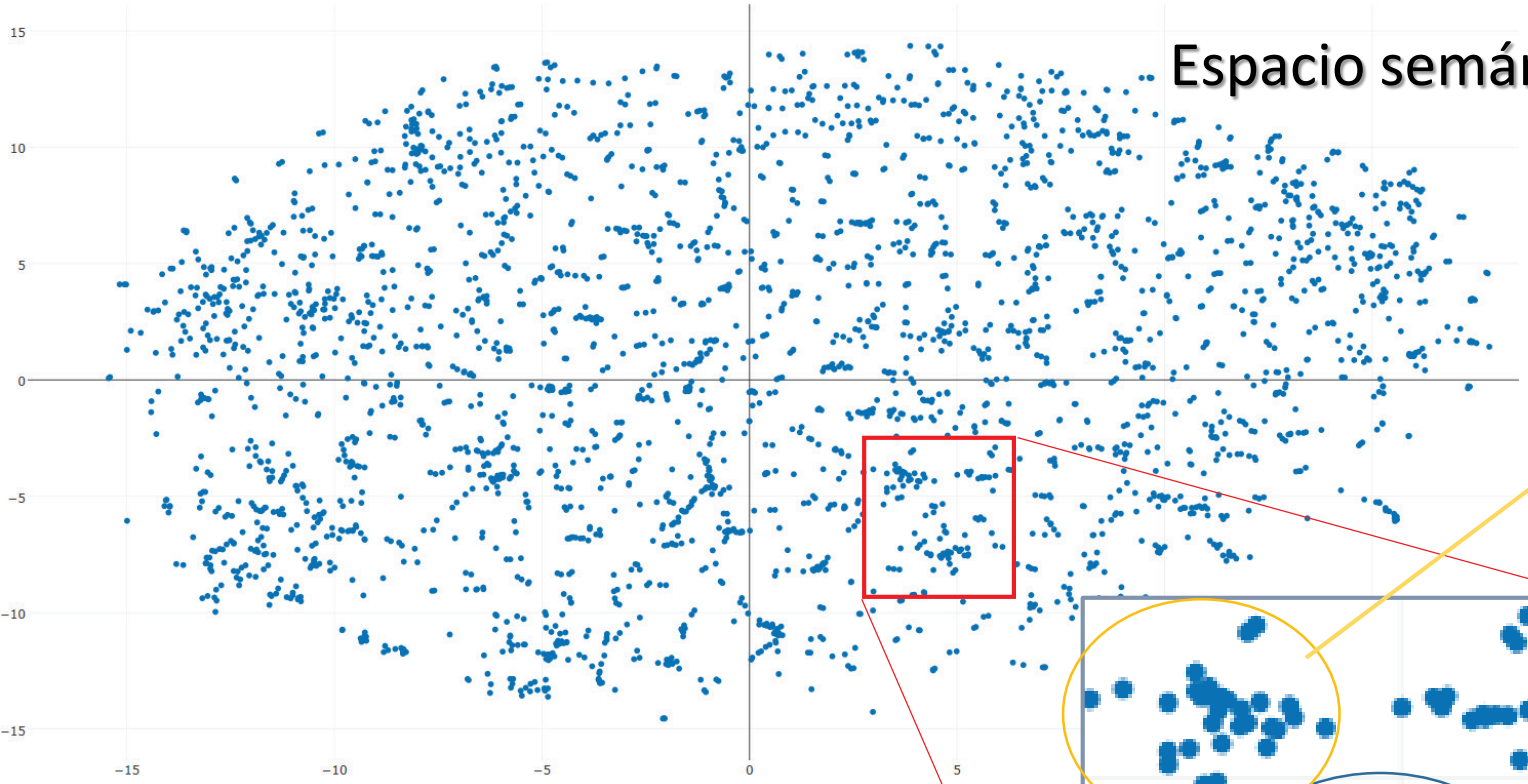
Dime con quién andas, y te diré quién eres

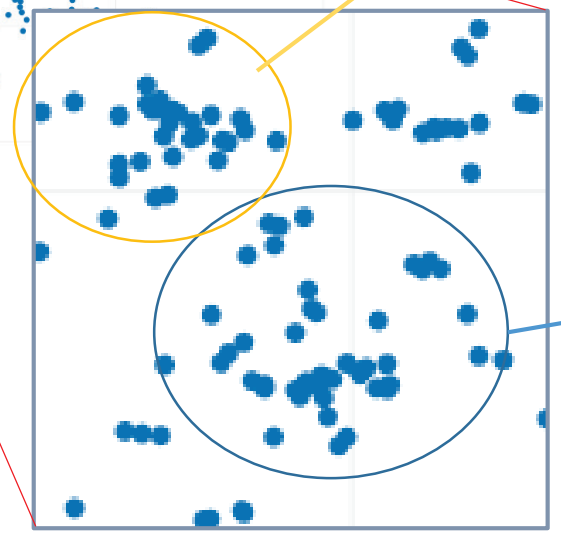
Hay una botella de *Belikin* sobre la mesa
A todo el mundo le gusta la *Belikin*
No bebas *Belikin* si tienes que conducir
La *Belikin* se fabrica con granos de cebada germinada

¿qué podemos deducir sobre las *Belikin* ?
Miramos las palabras que acompañan
Buscamos la similitud semántica con otras palabras ya conocidas
… y deducimos que la *Belikin* debe ser una bebida similar a…

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Espacio semántico: *embbedings*





Sótano
Pasillo
Patio
Ático
Salón
Balcón
…

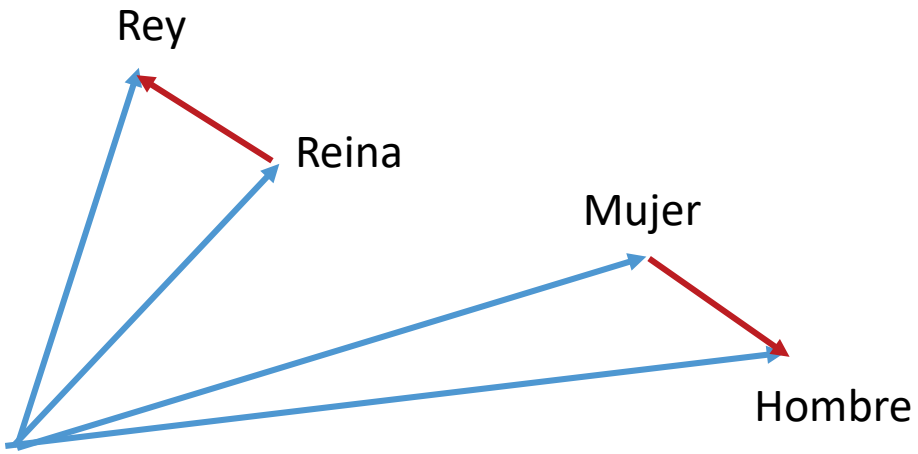
Fútbol
Béisbol
Rugby
Baloncesto
Tenis
Balonmano
Esquí
….

La proximidad semántica entre unidades lingüísticas se representa por la proximidad en el espacio

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Relaciones semánticas

vector[Reina] = vector[Rey] - vector[Hombre] + vector[Mujer]



Imágenes próximas

(Kiros, Salakhutdinov, Zemel, TACL 2015)

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Espacio semántico

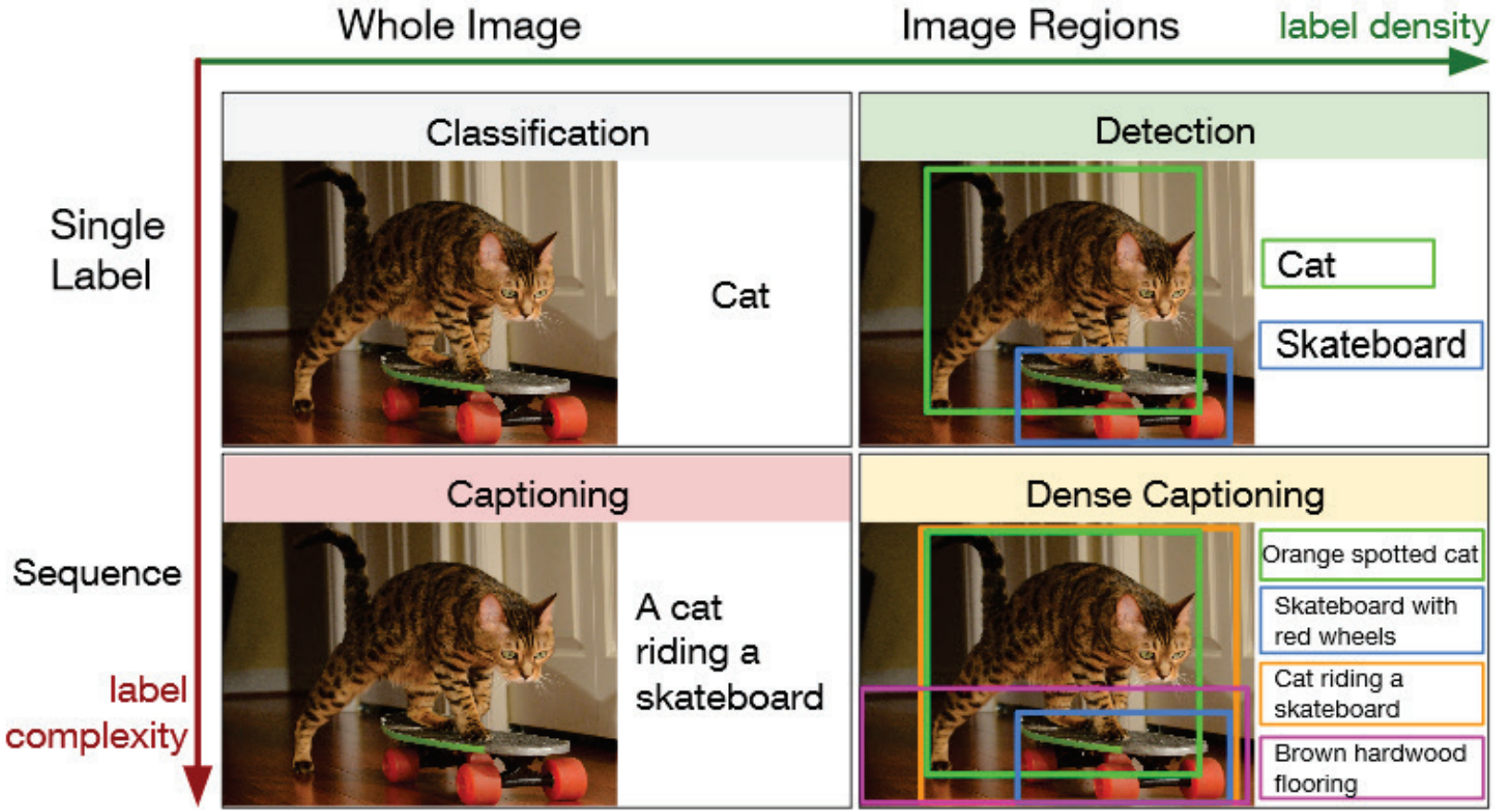
Video Summarization using Deep Semantic Features, Mayu Otani et al. 2016

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Espacio semántico: Generación de un resumen

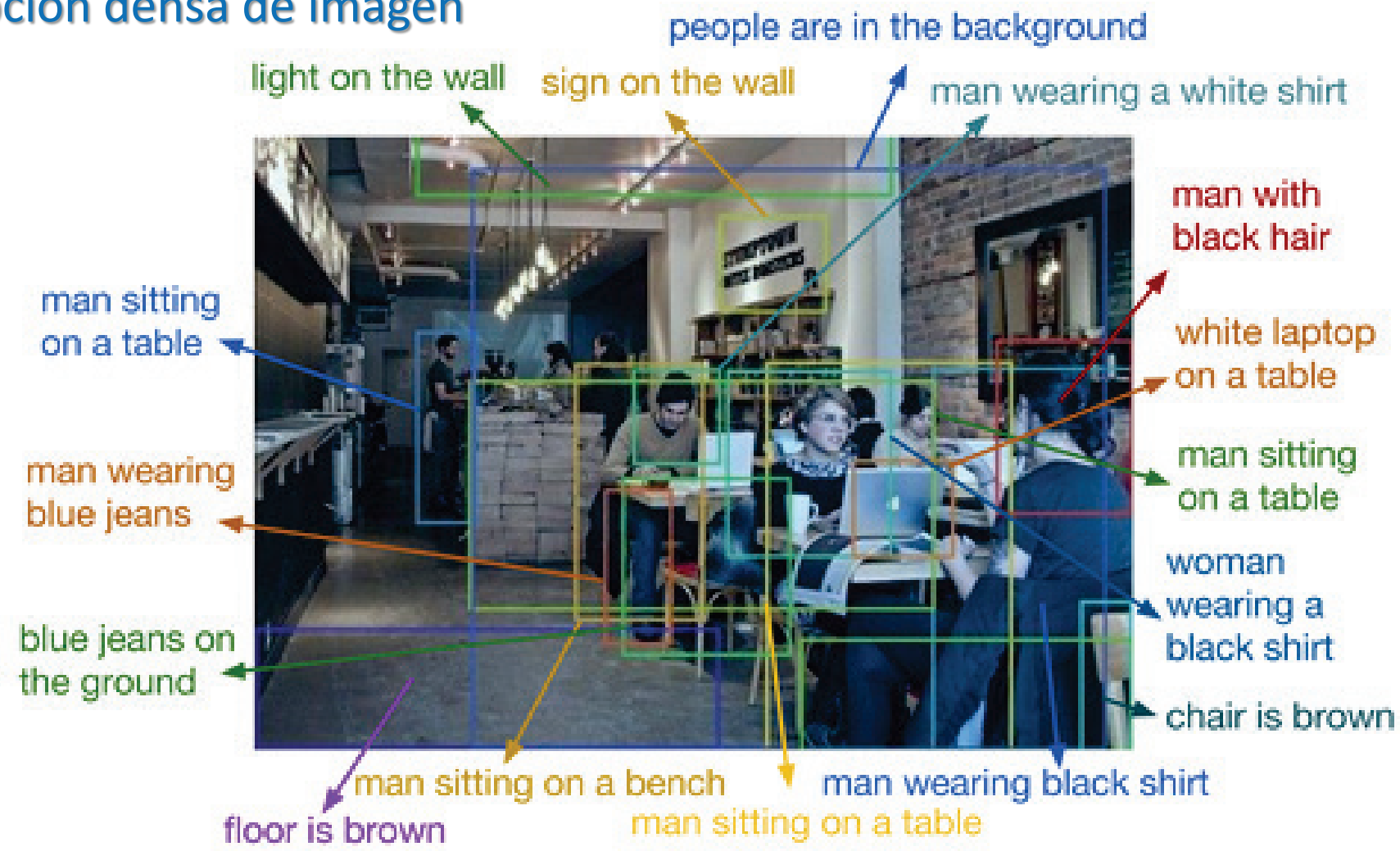
Video Summarization using Deep Semantic Features, Mayu Otani et al. 2016

Ejemplo de descripción de imagen



DenseCap: Fully Convolutional Localization Networks for Dense Captioning, Justin Johnson, Andrej Karpathy, Li Fei-Fei, Presented at CVPR 2016

Ejemplo de descripción densa de imagen



Tecnologías para el análisis y metadatado de contenidos audiovisuales

Ejemplo de descripción de imagen

A **caribou** that is laying in the grass.

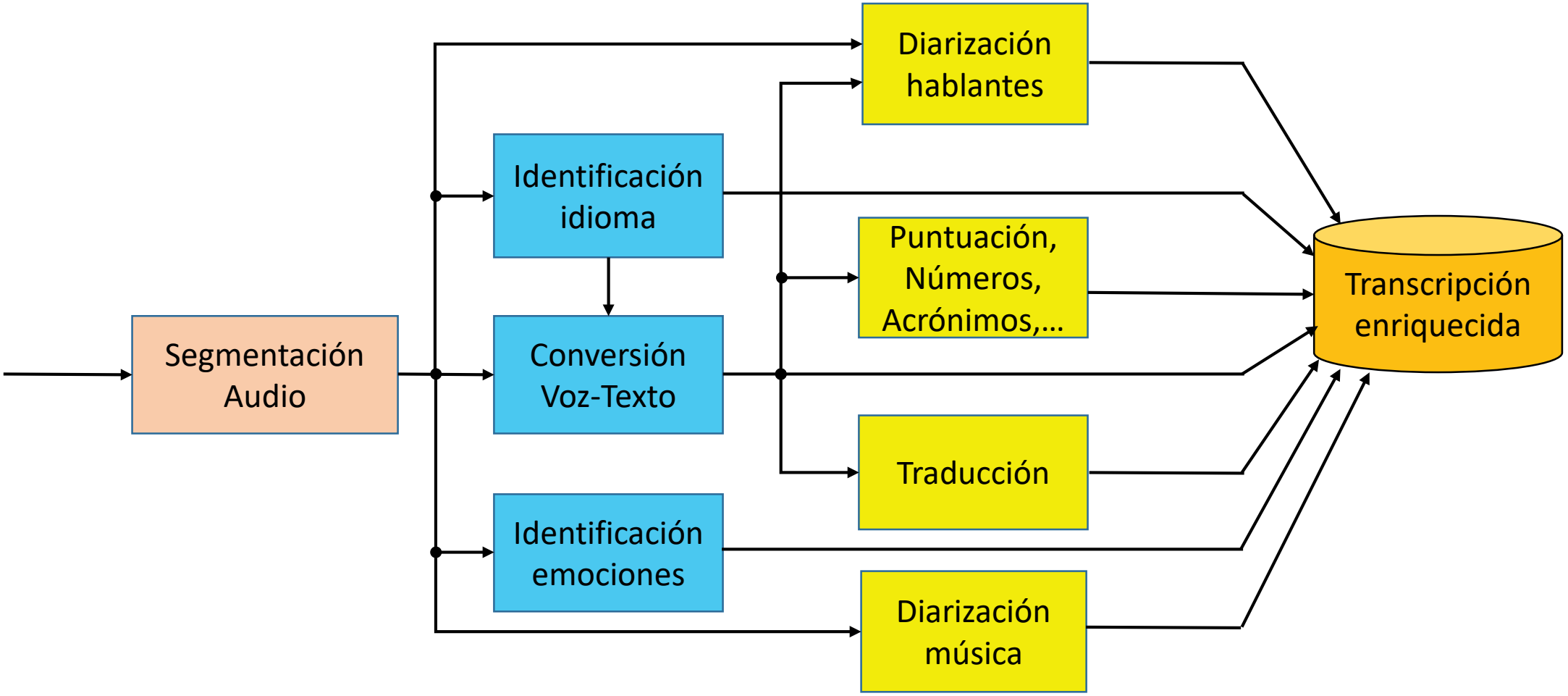
A large **warship** is on the water.

A man holding a **lychee** and lychee tree.

A **trampoline** with a **trampoline** in the middle of it.

Tecnologías para el análisis y metadatado de contenidos audiovisuales

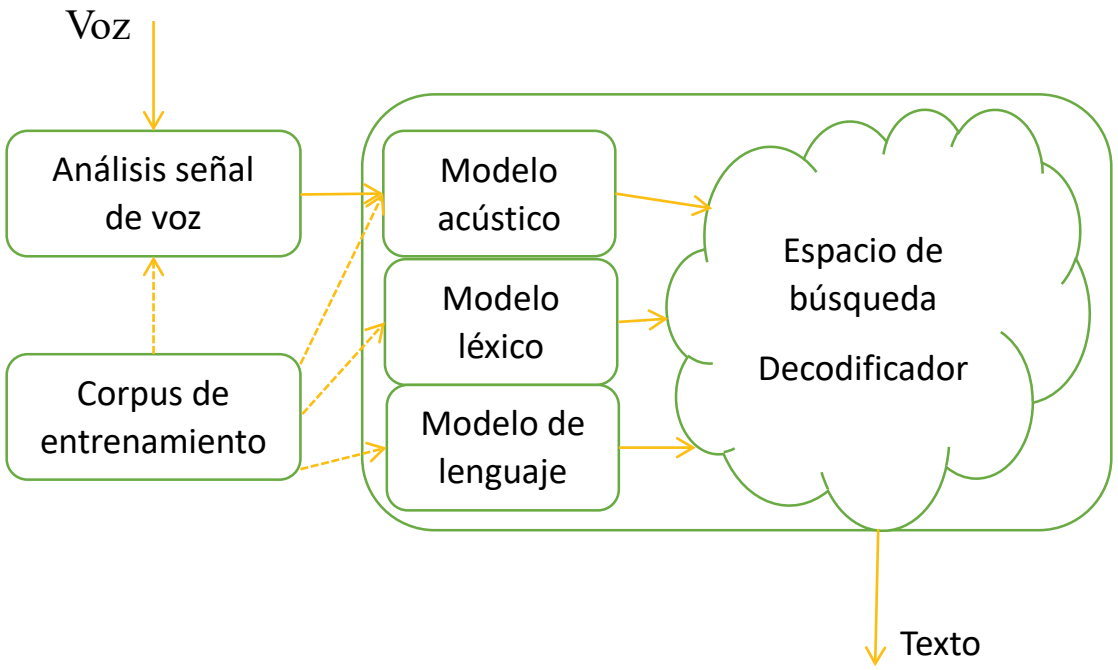
Ejemplo de aplicación de las tecnologías del habla



Tecnologías para el análisis y metadatado de contenidos audiovisuales

Ejemplo de aplicación de las tecnologías del habla

Conversión voz a texto



$$\text{tasa palabras correctas (Corr)} = \frac{\#\text{palabras correctas}}{\#total\ palabras\ reconocer} \times 100$$

$$\text{tasa palabras substituidas(Sub)} = \frac{\#\text{palabras substituidas}}{\#total\ palabras\ reconocer} \times 100$$

$$\text{tasa palabras borradas(Bor)} = \frac{\#\text{palabras borradas}}{\#total\ palabras\ reconocer} \times 100$$

$$\text{tasa palabras insertadas(Ins)} = \frac{\#\text{palabras insertadas}}{\#total\ palabras\ reconocer} \times 100$$

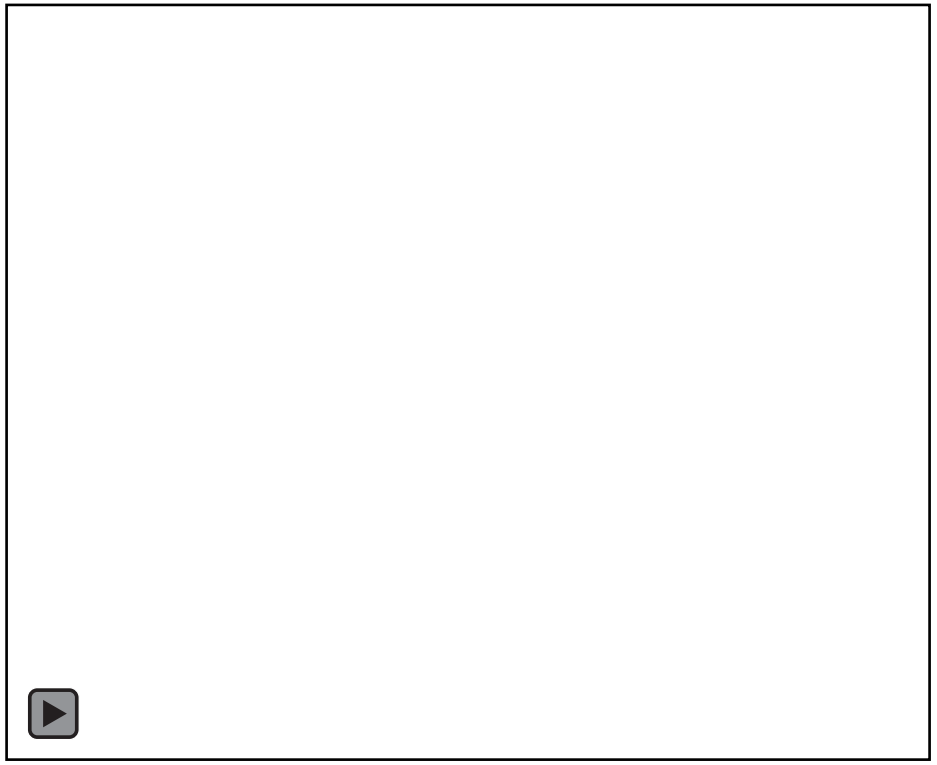
$$\text{tasa error palabras (Err)} = \frac{\#I + \#B + \#S}{\#total\ palabas\ reconocer} \times 100$$

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Conversión Voz a Texto

Programa noticiario 24H

Se trata de un noticiario completo del canal 24H, con una duración de unos 40 minutos.
Calidad de audio: muy buena en estudio, buena en la mayoría de los totales
Tipo de habla: leída en estudio y voz en off, espontánea en totales.



Noticiario 24H					
ESR_50_20171010_0009					
Programa	Corr	Sub	Borr	Ins	Error
Sin signos puntuación	93,7	3,6	2,7	1,1	7,5 (11,6)
Con puntos	92,4	4,3	3,3	2,6	10,2
Con puntos y comas	87,3	5,2	7,5	2,1	14,8

Locutor	frases	Corr	Sub	Bor	Ins	Err
Presentador	101	95.6	3.4	0.9	1.4	5.7
Nacho Lozano	11	95.6	3.7	0.7	3.0	7.4
Luis de Guindos	2	100	0.0	0.0	1.8	1.8
Rubén Urdiales	5	94.4	5.6	0.0	0.9	6.4
Rafael Catalá	2	89.0	7.3	3.7	8.5	19.5
Alfonso Guerra	3	87.5	10.4	2.1	6.3	18.8
Voz_off1	8	99.2	0.8	0.0	0.8	1.7
Voz_off2	6	96.8	3.2	0.0	1.3	4.5

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Conversión Voz a Texto Ejemplo de errores cometidos

Pares de confusión	Inserciones	Sustituciones	Borrados
5 -> sí ==> si	16 -> a	10 -> puigdemont	16 -> a
3 -> del ==> el	11 -> de	10 -> trump	8 -> y
3 -> piqué ==> que	10 -> y	9 -> sí	7 -> el
3 -> puigdemont ==> pulmón	6 -> el	7 -> a	5 -> en
3 -> trump ==> tram	5 -> en	6 -> de	5 -> ha
2 -> a ==> cuando	5 -> no	6 -> el	4 -> de
2 -> al ==> el	4 -> lo	5 -> y	4 -> que
2 -> carles ==> carlos	4 -> se	4 -> carles	3 -> esto
2 -> continua ==> continúa	3 -> es	4 -> ha	3 -> he
2 -> d'esquadra ==> esquadra	3 -> la	4 -> la	3 -> lo
2 -> dónde ==> donde	3 -> que	4 -> may	2 -> con
2 -> el ==> al	2 -> d	4 -> qué	2 -> jugando
2 -> hoy ==> y	2 -> hay	3 -> del	2 -> por
2 -> lluis ==> luis	2 -> más	3 -> en	2 -> se
2 -> omnium ==> unión	2 -> me	3 -> es	2 -> unos
2 -> por ==> porque	2 -> puse	3 -> hoy	1 -> analizar
2 -> porqué ==> porque		3 -> lluis	
2 -> qué ==> que		3 -> omnium	

Tecnologías para el análisis y metadatado de contenidos audiovisuales

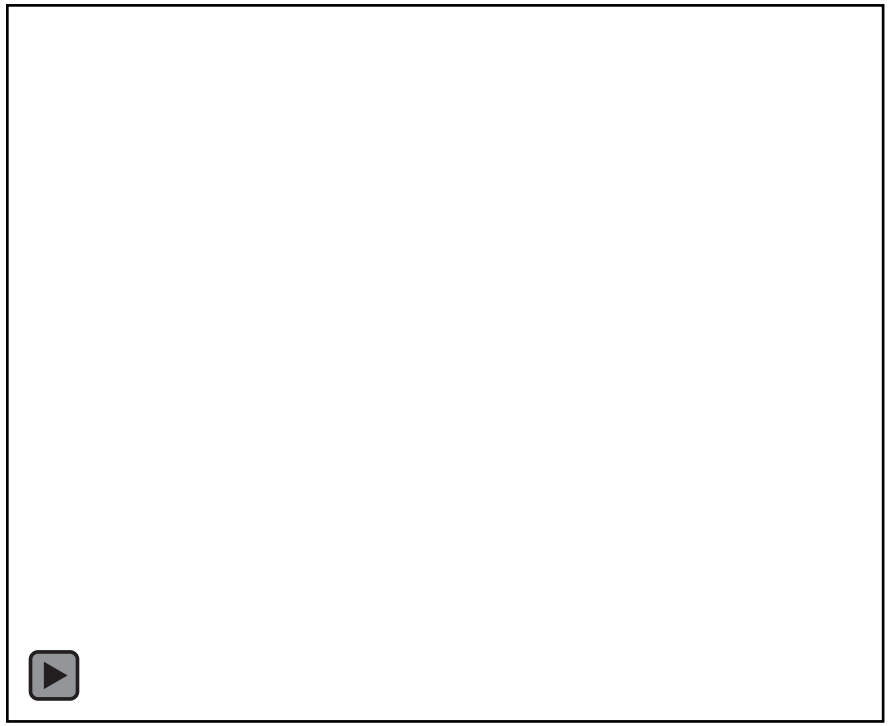
Conversión Voz a Texto

Programa Documental Histórico – San Saturio

Se trata de un documental en blanco y negro sobre las fiestas Toros en Soria

Calidad de audio: por lo general mala, hay música y voces en el fondo mientras se habla, zonas con solo música, voces con mucha reverberación. Hay una voz en off con un audio aceptable

Tipo de habla: mayoritariamente leída



Documental San Saturio					
Programa	Corr	Sub	Borr	Ins	Error
Sin signos puntuación	78.9	9.3	11.8	1.8	22.9 (27,5)
Con puntos	77.8	9.3	12.9	2.7	24.9
Con puntos y comas	72.3	9.6	18.1	2.0	29.7

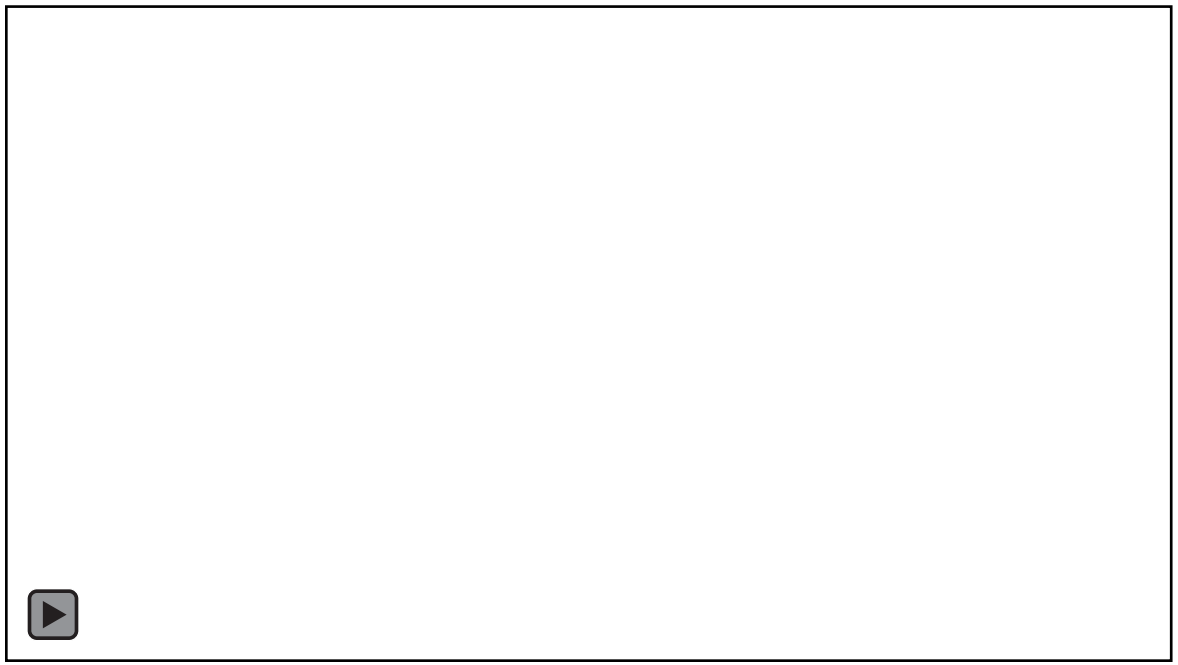
Locutor	frases	Corr	Sub	Bor	Ins	Err
Voz_off1	6	84.6	13.0	2.5	3.1	18.5
Voz_off2	53	91.6	6.3	2.2	1.4	9.8
C. J. Cela	7	54.2	17.3	28.5	2.5	48.4
Emilio Ruiz	15	86.3	10.2	3.4	3.7	17.3
J.A. Gaya Nuño	5	70.4	20.4	9.2	9.2	38.8

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Conversión Voz a Texto

Programa Variedades – Torres y Reyes

Se trata de un programa de variedades donde existen entrevistas, segmentos musicales, situaciones cómicas.
Calidad de audio: muy buena cuando solo habla una persona. Por lo general hay mucho solape de voces y música.
Tipo de habla: espontánea. Muchas inflexiones en la voz de uno de los presentadores.



Programa Torres y Reyes 12					
Programa	Corr	Sub	Borr	Ins	Error
Sin signos puntuación	76,4	8,2	15,5	1,6	25,3 (29,3)
Con puntos	72,3	12,5	15,2	4,8	32,5
Con puntos y comas	67,5	12,8	19,7	4,0	36,5

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Conversión Voz a Texto

Programa Informativo/Variedades – La Mañana

Se trata de un programa de información matinal con entrevistas, debates y directos.
Calidad de audio: buena
Tipo de habla: mayoritariamente espontánea.

Programa La Mañana					
Programa	Corr	Sub	Borr	Ins	Error
Sin signos puntuación	80,1	8,4	11,5	2,0	21,9 (26,5)
Con puntos	78,0	8,7	13,3	4,2	26,2
Con puntos y comas	71,2	10,8	18,0	4,1	32,9

Programa Musical – OT Gala 1

Se trata de un programa musical donde se promocionan nuevos cantantes
Calidad de audio: buena. Hay mucho solape y música. Muchos aplausos y gritos.
Tipo de habla: espontánea.

Programa OT					
Programa	Corr	Sub	Borr	Ins	Error
Sin signos puntuación, sin canciones	61,0	9,9	29,1	1,4	40,4 (47,7)

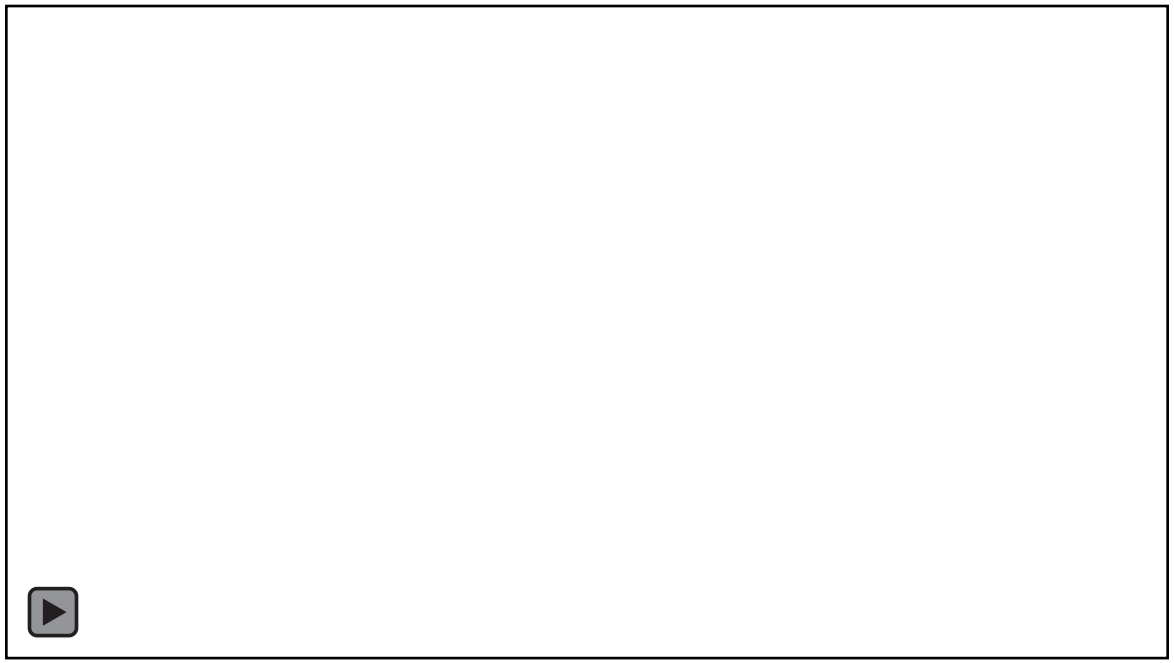
Programa Deportivo – Partido Fútbol

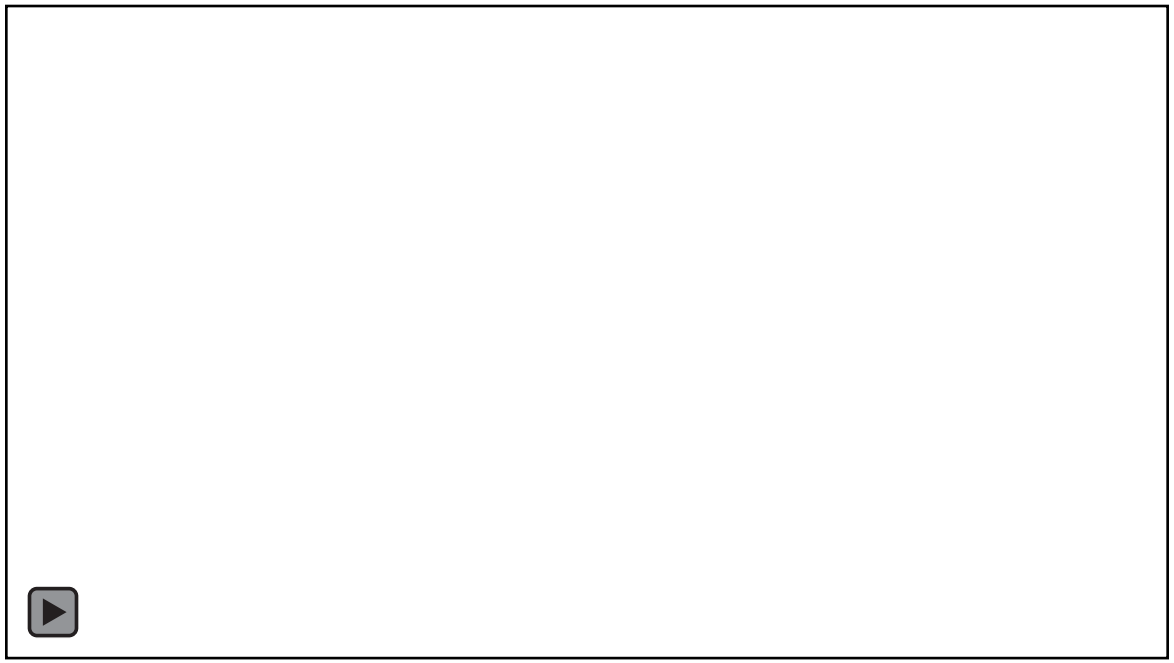
Se trata de una retransmisión en directo del partido España-Liechtenstein.
Calidad de audio: buena para el locutor principal. Reverberación en el resto. Aplausos y gritos que se solapan con el audio con un nivel considerable.
Tipo de habla: espontánea

Programa Deportes Futbol					
Programa	Corr	Sub	Borr	Ins	Error
Sin signos puntuación	57,3	12,0	30,7	3,0	45,7 (53,1)

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Conversión Voz a Texto





Tecnologías para el análisis y metadatado de contenidos audiovisuales

Conversión Voz a Texto

Programa La Mañana

Pares de confusión	Inserciones	Borrados	Substituciones
14 -> qué ==> que	45 -> que	65 -> a	38 -> de
11 -> sí ==> si	43 -> de	56 -> no	38 -> el
7 -> cómo ==>	32 -> la	51 -> de	34 -> a
8 -> solo ==>	31 -> a	52 -> que	33 -> sí
6 -> el ==>	30 -> no	50 -> y	30 -> que
7 -> no ==> nos	25 -> en	44 -> sí	31 -> y
8 -> si ==> y	20 -> el	43 -> es	27 -> es
5 -> ahí ==> y	21 -> y	42 -> en	26 -> no
6 -> de ==> del	14 -> es	32 -> el	25 -> lo
7 -> del ==> de	12 -> las	28 -> lo	24 -> la
8 -> el ==> de	13 -> lo	26 -> la	20 -> me
9 -> el ==> del	11 -> se	20 -> un	21 -> qué
10 -> ese ==> este	8 -> los	18 -> le	19 -> si
11 -> las ==> la	9 -> o	19 -> se	16 -> naiara
12 -> por ==> porque	10 -> sí	16 -> o	15 -> en 1
5 -> y ==> de	6 -> al	16 -> pero	15 -> ese

Tecnologías para el análisis y metadatado de contenidos audiovisuales

Conclusiones

- ✓ Grandes avances en los últimos 5 años en visión artificial, tecnologías del habla y procesamiento del lenguaje natural
  - Acceso a datos (Big Data)
  - Aprendizaje automático
  - GPUs
  - Software de código abierto

- ✓ Grandes esperanzas en lo que puede aportar la tecnología

¿Cómo podemos aprovechar lo mejor de las herramientas automatizadas y la experiencia humana?