Natural Language Generation

...

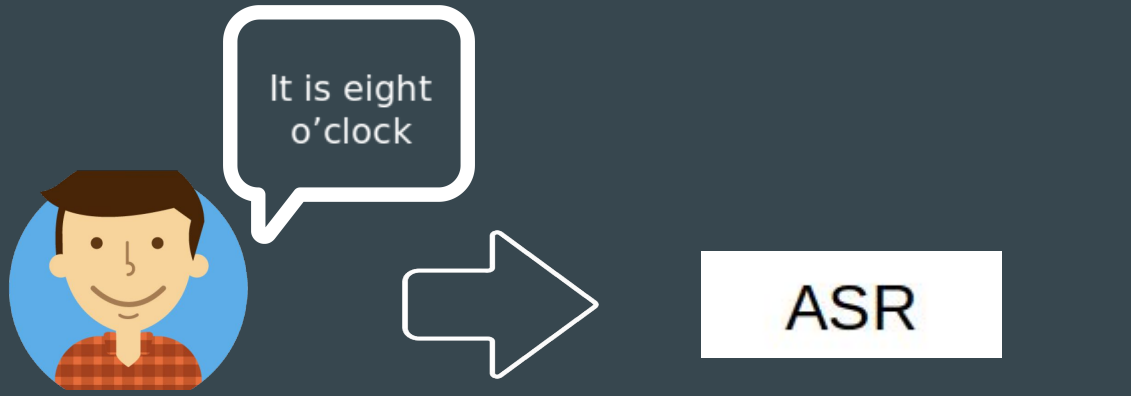
KeepCoding - Bootcamp de Big Data & Machine Learning

Índice

1. Modelos de Lenguaje

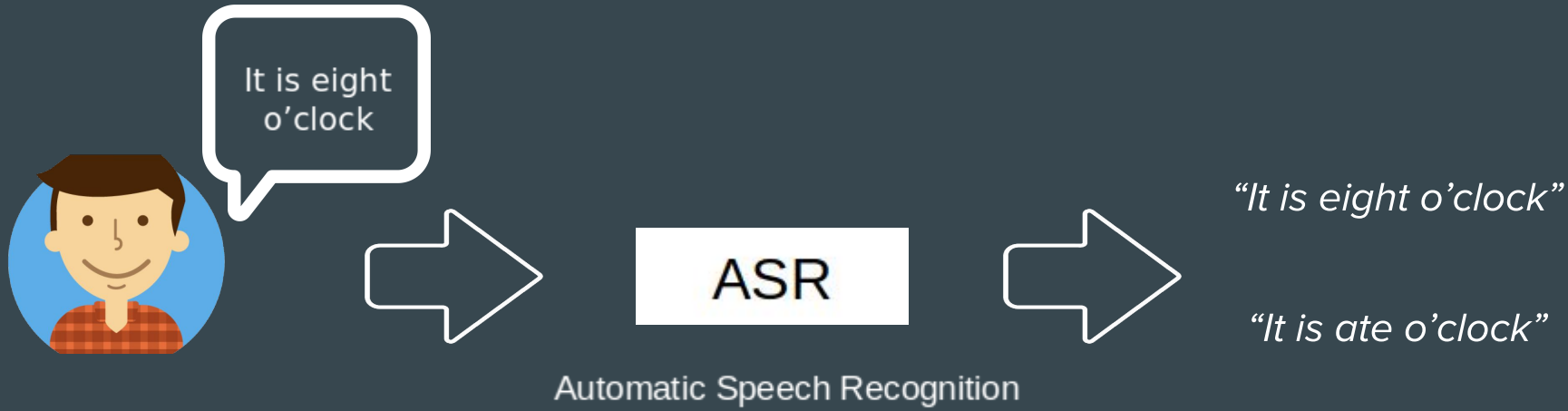
1. Modelos de Lenguaje

1. Modelos de Lenguaje

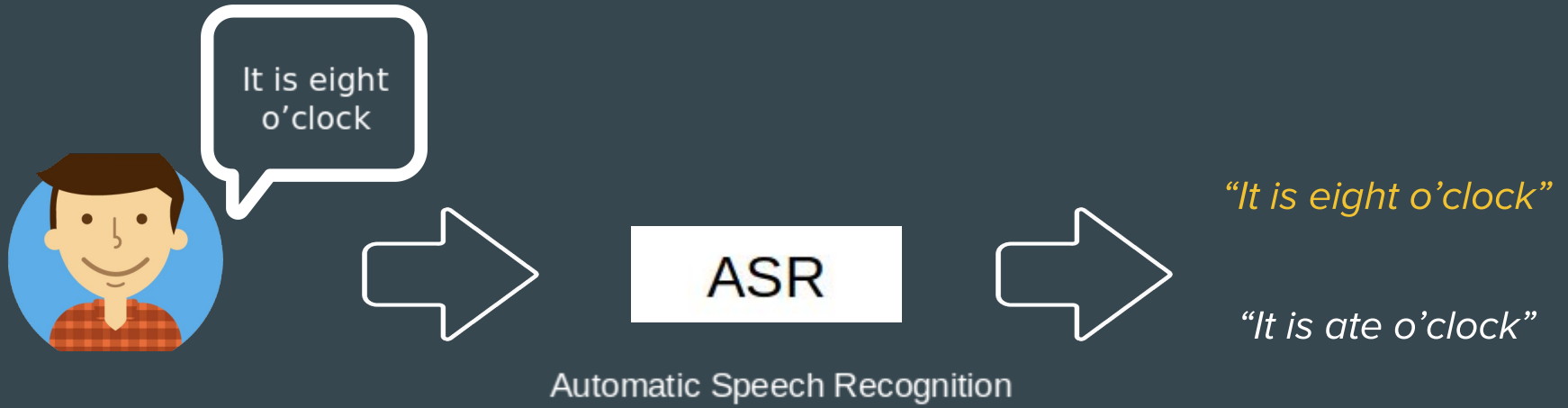


Automatic Speech Recognition

1. Modelos de Lenguaje



1. Modelos de Lenguaje



2. Modelos de Lenguaje - ¿Qué son?

Un modelo de lenguaje es un modelo estadístico que permite asignar una cierta probabilidad a una determinada secuencia de tokens mediante una distribución de probabilidad.

El objetivo es representar secuencias válidas de tokens por encima de aquellas que son incorrectas (siempre teniendo en cuenta el contexto).

P(“It is eight o’clock”) > P(“It is ate o’clock)

3. Modelos de Lenguaje - Etapas

1. Elección de un corpus (o conjunto de frases / documentos) que serán los datos en el entrenamiento del modelo. El idioma o el dominio (artículos científicos, chats en Internet, libros, etc.) son factores clave a tener en cuenta
2. Procesado de texto
    a. Normalizar las palabras
    b. Detectar y eliminar posibles errores gramaticales
    c. Conocer y entender mejor los datos con los que trabajemos
3. Tokenización de los documentos en frases, palabras o caracteres
4. Entrenamiento, validación del modelo y conclusiones

4. Modelos a nivel de carácter / palabra

Modelos de lenguaje a nivel de palabra:

1 Token = 1 Palabra

Modelos de lenguaje a nivel de carácter:

1 Token = 1 carácter

4. Modelos a nivel de carácter / palabra

- Coste computacional
  - Los modelos a nivel de carácter son más exigentes que los de nivel de palabra

4. Modelos a nivel de carácter / palabra

- Coste computacional
  - Los modelos a nivel de carácter son más exigentes que los de nivel de palabra
- Secuencias muy largas
  - Los modelos a nivel de palabra captan mejor las dependencias entre tokens lejanos en secuencias largas

4. Modelos a nivel de carácter / palabra

- Coste computacional
  - Los modelos a nivel de carácter son más exigentes que los de nivel de palabra
- Secuencias muy largas
  - Los modelos a nivel de palabra captan mejor las dependencias entre tokens lejanos en secuencias largas
- Vocabulario
  - Lo modelos a nivel de palabra tienen diccionarios muchísimo más grandes

4. Modelos a nivel de carácter / palabra

- Coste computacional
  - Los modelos a nivel de carácter son más exigentes que los de nivel de palabra
- Secuencias muy largas
  - Los modelos a nivel de palabra captan mejor las dependencias entre tokens lejanos en secuencias largas
- Vocabulario
  - Lo modelos a nivel de palabra tienen diccionarios muchísimo más grandes
- Palabras fuera del vocabulario (OOV)
  - Los modelos a nivel de carácter pueden asignar una probabilidad no nula a OOV

4. Modelos a nivel de carácter / palabra

- Coste computacional
  - Los modelos a nivel de carácter son más exigentes que los de nivel de palabra
- Secuencias muy largas
  - Los modelos a nivel de palabra captan mejor las dependencias entre tokens lejanos en secuencias largas
- Vocabulario
  - Lo modelos a nivel de palabra tienen diccionarios muchísimo más grandes
- Palabras fuera del vocabulario (OOV)
  - Los modelos a nivel de carácter pueden asignar una probabilidad no nula a OOV
- Puntuación
  - Los modelos a nivel de palabra tienen muy difícil contemplar todas las posibilidades

5. Retos

- Disponibilidad de un corpus
- Errores gramaticales
- Riesgo de perder información durante el preprocesado
- Palabras fuera del vocabulario (OOV words)
- Alta cardinalidad del vocabulario
- Potencia de cómputo
- Evaluación

6. Ejemplos

Corpus 1

Teclado predictivo

- '*Más allá del bien y del mal*', Nietzsche (1886)
- 600.000 caracteres aprox
- 57 caracteres distintos

Corpus 2

Generación de poesía automática

- *Sonetos*, Shakespeare (1609)
- 100.000 caracteres aprox
- 48 caracteres distintos

Preprocesado (para ambos corpus)

- Lowercase, chunks de 40 caracteres (el 41 para predecir), saltos de 3 caracteres, vectores one-hot-encoding.

Modelo (para ambos corpus)

- Diferentes LSTM + Capas densas

Train / Test (para ambos corpus)

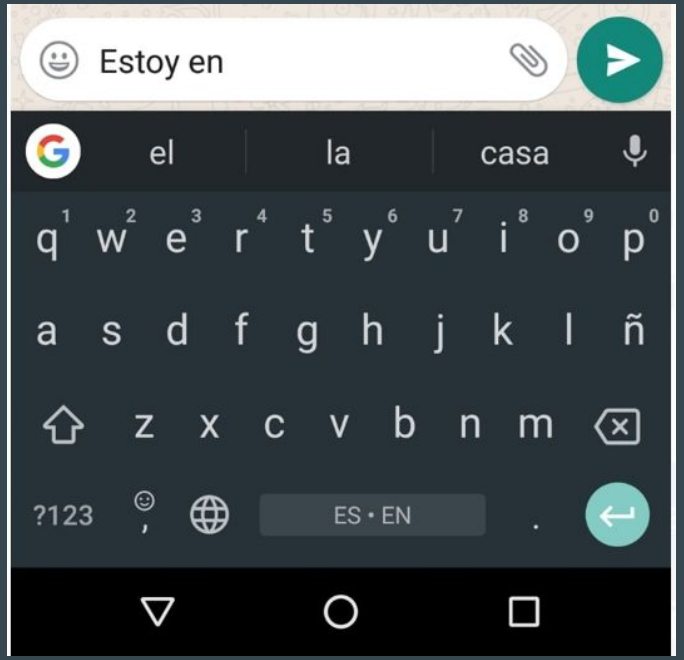
- 5 % para validación
- Accuracy como figura de mérito
- Categorical Cross-Entropy como loss function

6. Ejemplos - Teclado Predictivo

Conforme escribimos el teclado de nuestro dispositivo trata de predecir las siguientes palabras (o caracteres).

Dicha predicción se realiza teniendo en cuenta lo escrito hasta ese momento.

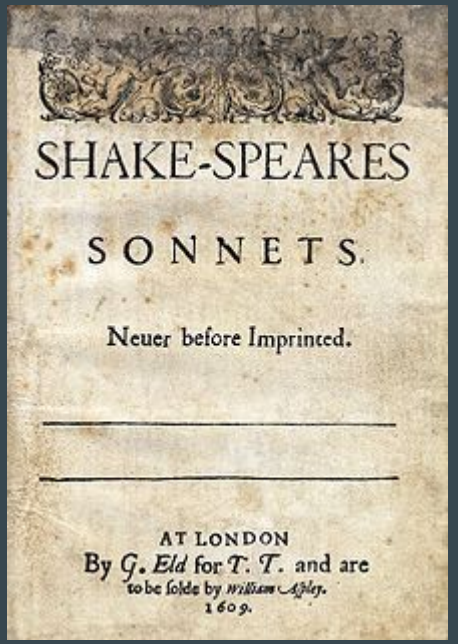
Además, puede incluir la corrección de posibles errores ortográficos.



6. Ejemplos - Escritura Automática de Poemas

La generación del lenguaje puede emplearse no solo en tratar de predecir el siguiente carácter o palabra si no en, por ejemplo, redactar noticias, crear respuestas en un chatbot o crear nuevo contenido copiando el estilo de un/a autor/a.

Sistemas aún más complejos de generación de poemas pueden ser entrenados usando también la representación fonética de los caracteres.



6. Ejemplos - Escritura Automática de Poemas (bonus)

sometimes alone
followed by vistas
you are a light seeker
and the light finds you

come on down to my boat baby
come on down where we can play
come on down to my boat baby
come on down we'll sail away

do colours really convey moods
colors answer feeling in man
shapes answer thought
motion answers will

the man bent over his guitar
a shearsman of sorts the day was green
they said 'you have a blue guitar
you do not play things as they are

in the celtic tongue
a glen is any dale
touched by the natural magic
of green shade

my walls outside must have some flowers
my walls within must have some books
a house that's small a garden large
and in it leafy nooks

what is life
it is the flash of a firefly in the night
it is the breath of a buffalo in the wintertime
it is the little shadow which runs across the grass and
loses itself in the sunset

but now the psyche of thy being
still shyly doth essay her delicate wing
like to that airy nurseling of the sun
when first it breaketh through its dun

we are all falling this hand is falling too
all have this falling sickness none withstands
and still there's always one whose gentle hands
this universal falling can't fall through

is it so small a thing
to have enjoy'd the sun
to have lived light in the spring
to have loved to have thought to have done

¡Vamos al lío!