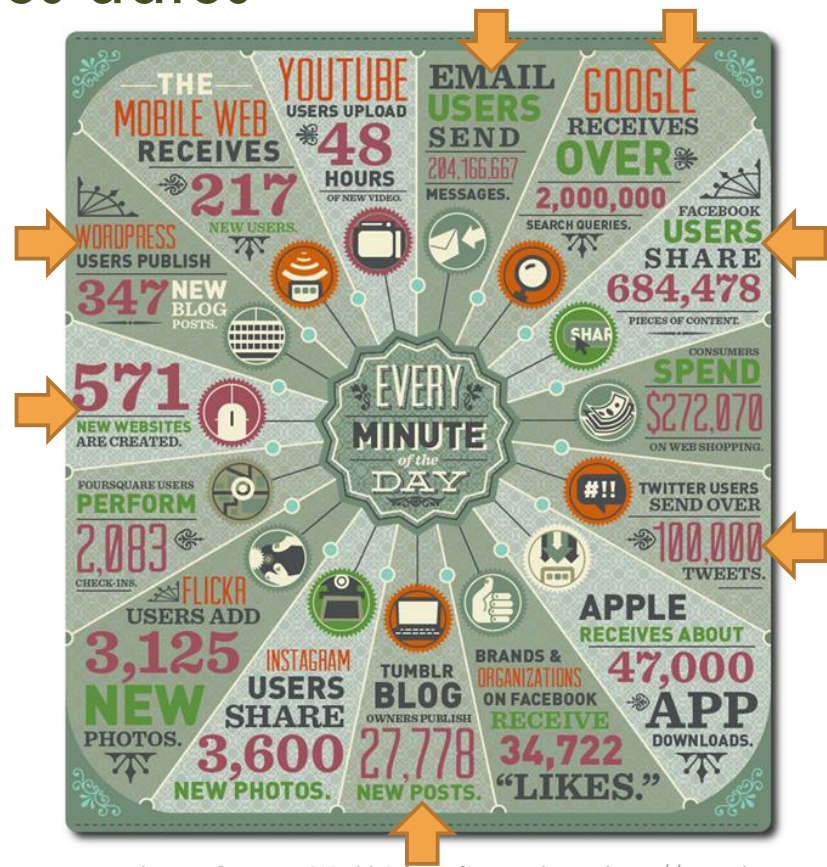


Análisis de textos

Introducción

El actual volumen de texto en los datos

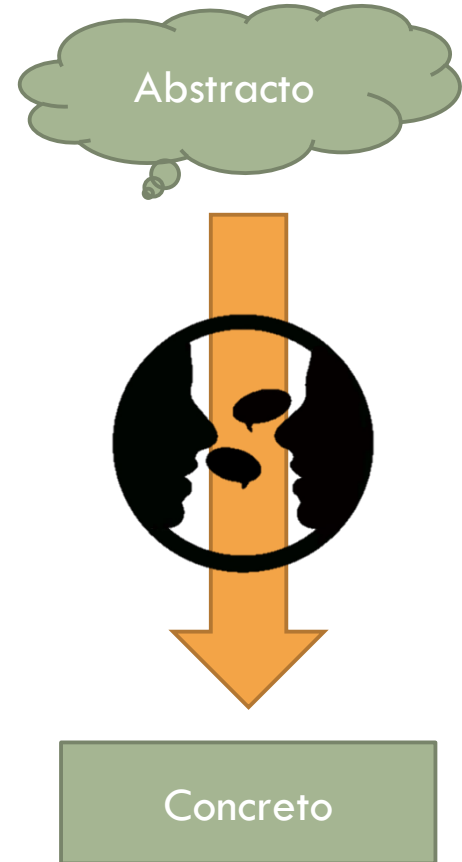
- Se estima que el 70%-80% de los datos que se generan son no estructurados
 - Gran parte de estos son en forma de **texto libre** o lenguaje natural
- Esta información no es procesable si no es mediante técnicas especialmente diseñadas para el texto



Estimación de % datos no estructurados por Computer World. Infografía creada por <https://www.domo.com/>

Lenguas naturales

- **Lenguaje natural:** capacidad cognitiva que permite al ser humano comunicarse
- **Lengua natural:** alguna de las variedades particulares del lenguaje natural
 - Conjunto finito de símbolos, pero permite la formación de un conjunto potencialmente infinito de enunciados
 - Distinto para cada hablante
 - Informal, no permite un procesamiento sistemático de su estructura y significado
- **Lengua estándar:** estandarización artificial del habla de un conjunto de hablantes
 - Formalización de la lengua que permite su estudio
- **Ortografía:** conjunto de reglas y convenciones que rigen el sistema de escritura habitual establecido para una lengua estándar



Lingüística y computación

➤ Ciencias de la computación

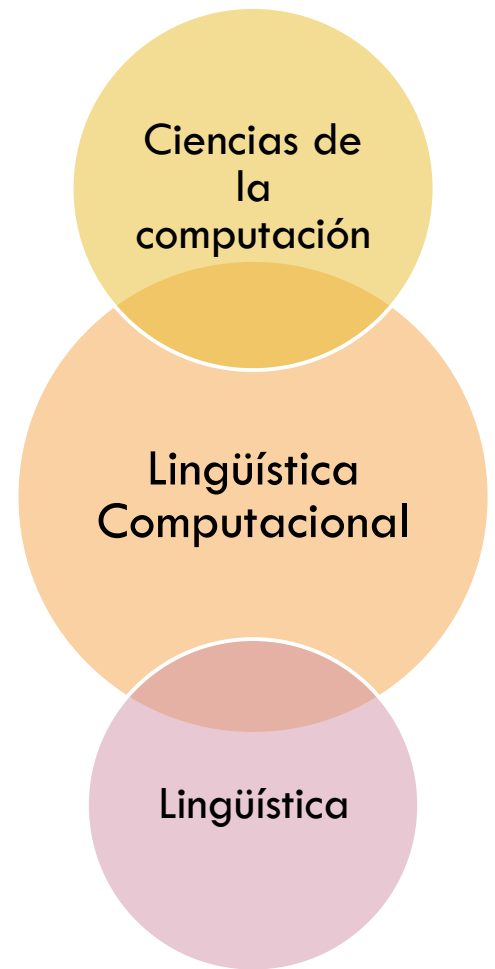
- Estudio sistemático de los **procesos algorítmicos** que describen y transforman **información**: su teoría, análisis, diseño, eficiencia, implementación y aplicación

➤ Lingüística

- Estudio científico tanto de la estructura de las **lenguas naturales** y de aspectos relacionados con ellas como de su evolución histórica, de su estructura interna y del conocimiento que los hablantes poseen de su propia lengua

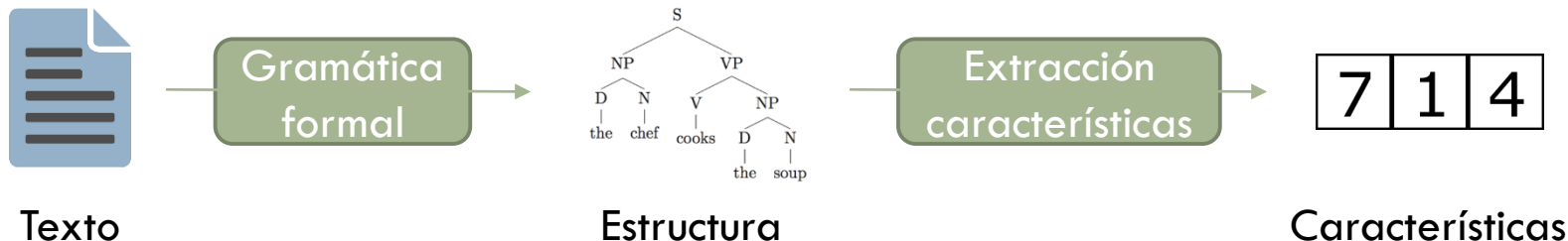
➤ Lingüística computacional

- Modelado basado en reglas y el modelado estadístico de las **lenguas naturales** desde una **perspectiva computacional**



Lingüística computacional para procesamiento de datos

- El lenguaje natural es, por naturaleza, **informal**
 - No permite un procesamiento sistemático de su estructura y significado
- Para tratar adecuadamente el texto necesitamos de **formalismos** que nos permitan aproximarnos al lenguaje natural de una forma disciplinada y no ambigua
- La **lingüística computacional** se basa en definir **gramáticas formales**, **diccionarios**, **ontologías** y otros recursos que nos permiten realizar estos análisis, transformando el lenguaje en estructuras bien definidas que podemos procesar automáticamente



Dificultades en el análisis de textos: ambigüedad

Time flies like an arrow

- El tiempo vuela como una flecha
- Las moscas del tiempo, como una flecha
- A las moscas del tiempo les gusta una flecha
- Cronometra moscas como lo haría una flecha
- Cronometra moscas como lo harías con una flecha
- Cronometra a las moscas que son como una flecha
- La revista Time vuela como una flecha

Dificultades: ironía

- Salió de la cárcel con tanta honra, que le acompañaron doscientos cardenales; salvo que a ninguno llamaban eminencia. (F. de Quevedo, Buscón)
- ¿Vendrás por la tarde a mi casa?
- Sí, pensaba pasar la tarde en el circo
- Me compraría el Ferrari ahora mismo, lo que pasa es que no tengo sueldo

Dificultades: el desiderátum del lenguaje



Cerrar podrá mis ojos la postrera
Sombra que me llevare el blanco día,
Y podrá desatar esta alma mía
Hora a su afán ansioso lisonjera;

Mas no, de esotra parte en la ribera,
Dejará la memoria en donde ardía:
Nadar sabe mi llama el agua fría,
Y perder el respeto a ley severa.

Alma a quien todo un dios prisión ha sido,
Venas que humor a tanto fuego han dado,
Médulas que han gloriosamente ardido,

Su cuerpo dejarán no su cuidado;
Serán ceniza, mas tendrá sentido;
Polvo serán, mas polvo enamorado.

Francisco de Quevedo

Dificultades: realidad del lenguaje



Seguir

@calisixd los abes los konstrulle el hestado, renfe solo pone trenes en la bia, rekuer2 a tu hamijo ;-))

Responder Retwittear Favorito Más

RETWEETS
6 278

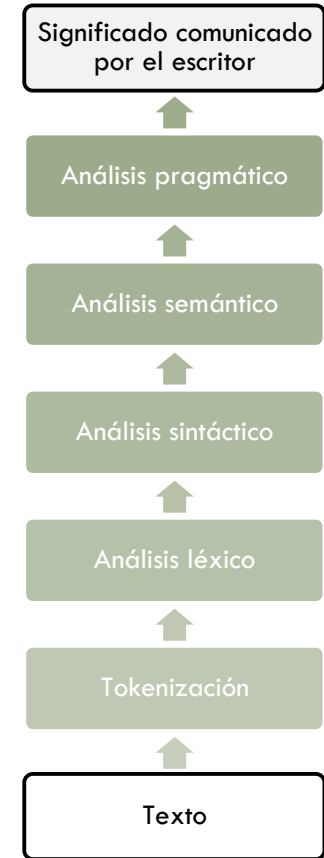
FAVORITOS
3 384



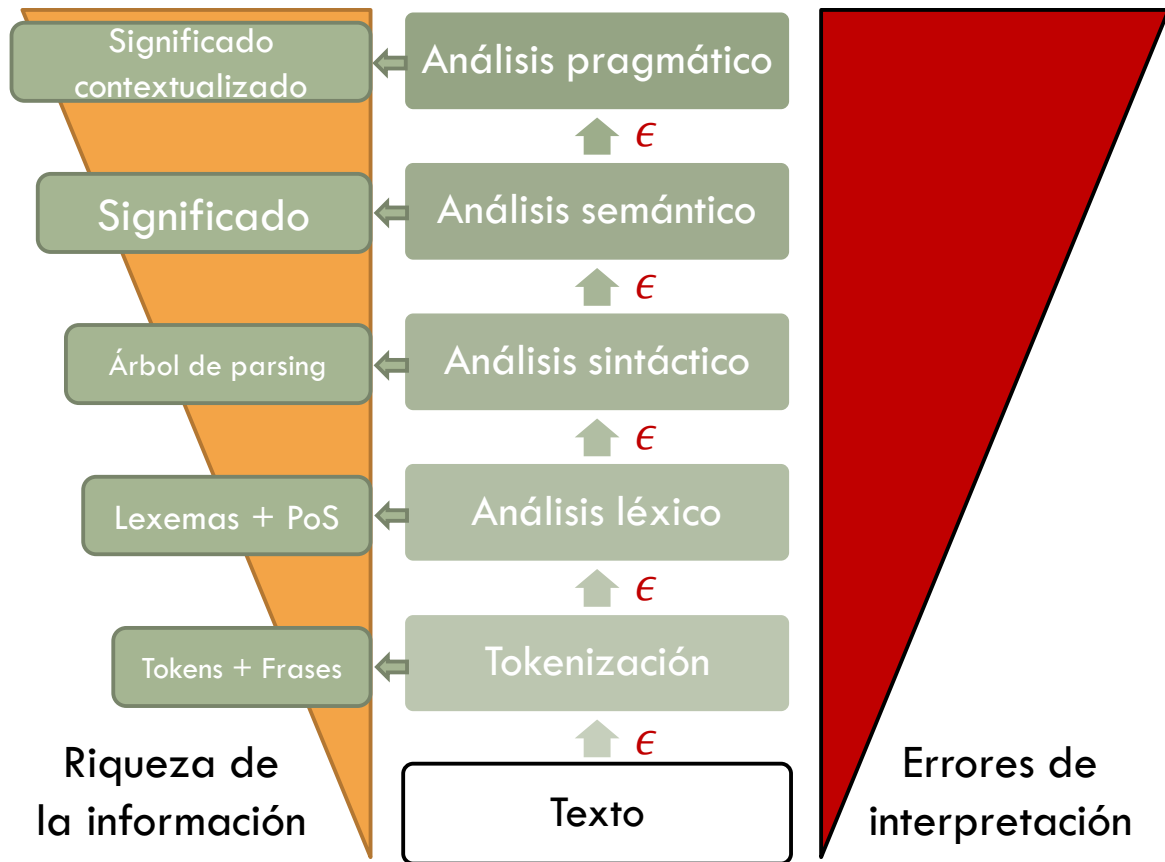
17:12 - 24 de abr. de 2014

Estratos de procesamiento

- El lenguaje natural en texto está formado por una serie de estratos de información, que pueden exponerse mediante procesos de análisis
 - **Tokenización**: segmentación en frases, palabras y caracteres
 - **Análisis léxico**: morfología, estructura interna de las palabras, declinaciones, ...
 - **Análisis sintáctico**: sintaxis, función de las palabras en sintagmas, oraciones y frases; relaciones entre palabras
 - **Análisis semántico**: significado de las palabras
 - **Análisis pragmático**: contexto y su influencia en el significado



Complejidad de análisis de los estratos lingüísticos



- A mayor profundidad de análisis, se obtiene mayor **riqueza de información** sobre el texto
- Pero los **errores** en los análisis y la interpretación del lenguaje son también mayores
- Los niveles inferiores de análisis están más desarrollados para procesamiento automático
 - El nivel pragmático solo se suele tener en cuenta para definir cómo enfocar el resto de análisis según la aplicación
- En técnicas modernas de análisis es habitual mezclar niveles: ej. léxico+sintáctico+semántico a la vez.
- El nivel de análisis requerido y de error aceptable **dependerá de la aplicación**

Corpus

- El elemento básico para hacer cualquier análisis en lingüística computacional es un corpus
- **Corpus** (plural corpora): conjunto amplio y estructurado de ejemplos reales de uso de la lengua (dataset no supervisado)
- **Corpus anotado**: corpus al que uno o varios expertos han añadido anotaciones de análisis de los textos (dataset supervisado)
 - Separaciones en palabras y frases (tokenización)
 - Morfología de las palabras (análisis léxico)
 - Árboles de parsing (análisis sintáctico)
 - Significado de las palabras (análisis semántico)
- Ejemplos de corpora
 - **Penn Treebank**: corpus en inglés anotado con morfología y parsing
<https://www.cis.upenn.edu/~treebank/>
 - **CREA** (Corpus de Referencia del Español Actual): corpus en español, no anotado
<http://www.rae.es/recursos/banco-de-datos/crea>
 - **Sentiment Treebank**: corpus en inglés anotado con parsing y semántica de la opinión
<http://nlp.stanford.edu/sentiment/treebank.html>
 - **Common Crawl**: corpus multilingüe no anotado, volcado de todo internet <http://commoncrawl.org/>

Estrategias de análisis

Estrategia clásica

Expertos lingüistas desarrollan **gramáticas y reglas del lenguaje**.

Se emplean **corpora** de texto anotados para **validar** las gramáticas y reglas.

Las gramáticas y reglas se explotan para analizar nuevos textos.

Estrategia híbrida

Se usan **gramáticas y reglas del lenguaje** para extraer **características complejas**.

Se emplean **métodos estadísticos** para entrenar modelos predictivos sobre corpora etiquetados.

Estrategia estadística

Se parte de **corpora** anotados con los objetivos que interesa predecir.

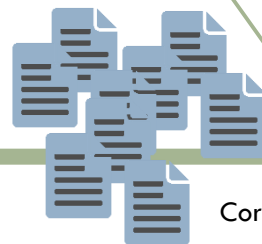
Se **extraen características básicas** del texto, y con **métodos estadísticos** (ej. Aprendizaje automático) se entrenan modelos predictivos.



Corpus anotado
pequeño



Corpus anotado
intermedio



Corpus anotado
grande



Álvaro Barbero Jiménez



@albarji



<https://github.com/albarji>



albarji.deviantart.com