Saturdays.AI
Murcia
MACHINE LEARNING

Introducción al NLP

Nicolás José Fernández Martínez
Profesor del Departamento de Idiomas (UCAM)

Tabla de contenidos

1. ¿Qué es el NLP?

- **Interdisciplinariedad** entre áreas como la Lingüística Computacional, Ciencia de Datos, y la Inteligencia Artificial.
- Centrado en el **diseño y análisis** de algoritmos y representaciones computacionales para tratar con **lenguas naturales** (discurso hablado y/o escrito) (Cambria & White, 2014).
- Objetivo: proporcionar soluciones computacionales a **problemas reales** que giran en torno al manejo de **datos textuales** (Periñán-Pascual, 2012), e.g. extracción de información, traducción automática, diálogo y conversación, ejecución de instrucciones, etc.
- A un dataset de datos textuales también se le llama **corpus**, o colección de textos.

1.1. Naturaleza del lenguaje humano

- **Composicionalidad** y **recursividad**: las palabras se combinan para formar sintagmas, los sintagmas se combinan para formar sintagmas más complejos (oraciones), las oraciones se combinan para formar párrafos, los párrafos se combinan para formar textos.
- El **significado** se construye a partir de esta combinación ordenada de unidades simbólicas (i.e. palabras o tokens), **pero no siempre** (Riemer, 2010).

1.1. Naturaleza del lenguaje

niño
el niño
el niño pequeño
el niño pequeño disfrazado de monstruo
el niño pequeño disfrazado de monstruo come
el niño pequeño disfrazado de monstruo come chocolate
el niño pequeño disfrazado de monstruo come chocolate en el patio

1.1. Naturaleza del lenguaje

- ¿Puede el significado derivarse de esta combinación de palabras?

estirar la pata
estar como una cabra
ser pan comido
ponerse como un tomate
no tener pies ni cabeza

Hace mucho calor
Ya veo que eres una persona muy ordenada

1.1. Naturaleza del lenguaje

- El lenguaje también tiene una **dimensión paradigmática**:

  La niña juega al fútbol
  La niña juega *al tenis*
  El niño juega *al baloncesto*
  El niño juega *un partido de baloncesto*
  El niño juega *con la pelota*
  El hombre juega *a dos bandas*
  El hombre *actúa* a dos bandas

1.2. Polémica y controversia: futuro del NLP y AI

- Noam Chomsky vs Peter Norvig
- Los modelos estadísticos de lenguaje natural podrían simular y predecir el comportamiento del lenguaje, pero no dicen nada acerca de su naturaleza, ni necesitan conocimiento lingüístico explícito integrado (poco rigor científico).
- Lenguaje natural: sistema cognitivo complejo (sentido común, razonamiento, inferencia)
- Gramaticalidad: probablemente, una cuestión de probabilidad
  *El grupo de albañiles **trabaja** hasta mediodía*
  *El grupo de albañiles **trabajan** hasta mediodía*

2. Tipos de enfoques (Periñán-Pascual, 2012)

- Enfoques simbólicos
- Enfoques probabilísticos: ML, DL

2.1. Enfoques simbólicos

- Basados en **reglas** de dominio específicas (e.g. expresiones regulares), aunadas con **conocimiento y recursos lingüísticos** (e.g. lexicones, diccionarios, ontologías, información gramatical, etc.)
- Muy útiles para la extracción de información: gran precisión, poca cobertura (Jurafsky & Martin, 2019)
- Enfoque muy utilizado en el mundo empresarial, en oposición al mundo académico (Chiticariu et al., 2013)

2.1. Enfoques simbólicos

- Ventajas:
  - No se necesitan datos de entrenamiento (i.e. training dataset) para su construcción y desarrollo
  - Eficiencia
  - Escalabilidad y reusabilidad

- Desventajas:
  - Exige conocimientos avanzados y específicos de la lengua
  - El desarrollo de las reglas es un proceso manual, costoso a nivel de tiempo

2.1. Enfoques probabilísticos

- Sistemas que no necesitan de forma explícita conocimiento lingüístico, sino que estiman **probabilidades**, a raíz de la información contextual proporcionada por la combinación de palabras, para **inferir** dicho **conocimiento**.
- Machine Learning: Hidden Markov Model, Conditional Random Field, Support Vector Machine, Naïve Bayes...
- Deep Learning: Convolutional Neural Networks, Recurrent Neural Networks, LSTM, Transformers...
- Parte de un conjunto de datos textuales manualmente etiquetado para su entrenamiento, normalmente extenso.

2.1. Enfoques probabilísticos

- Ventajas:
- No hay necesidad de tener un extenso conocimiento lingüístico
- Gran rendimiento en todo tipo de tareas de NLP

- Desventajas:
- Se necesitan gran cantidad de datos y recursos computacionales (e.g. Hardware) para el proceso de entrenamiento.
- A veces, sigue siendo necesario añadir características lingüísticas explícitas (i.e. *feature engineering*) para conseguir un mayor rendimiento.

3. Tipos de tareas

- Tokenización (tokenization)
- Lematización (lemmatization)
- Etiquetación gramatical (POS tagging)
- Análisis sintáctico (Syntactic parsing)
- Extracción de entidades nombradas (NER)
- Análisis de sentimiento o minería de opinión (Sentiment Analysis / Opinion mining)
- Similitud semántica con Word Embeddings
- Detección del tema (Topic Detection / Analysis)
- Resumidor (Text Summarization)
- Traductores automáticos (Machine Translation)
- Agentes conversacionales (Chatbots)

3.1. Tipos de tareas: tokenización

- La tarea más sencilla, implica la **segmentación** de trozos de texto en tokens o **palabras**.
- Normalmente un tokenizer trabaja mediante expresiones regulares que tienen en cuenta los espacios y signos de puntuación como delimitadores de palabras.
- Suele ser el primer paso en un *pipeline* de NLP.
- Cada lengua tiene su idiosincrasia: en español, pronombres clíticos en verbos (e.g. negarse); en inglés, formas contraídas (e.g. isn't), en chino los espacios no delimitan palabras.

3.1. Tipos de tareas: tokenización

```
from nltk.tokenize import word_tokenize
```

```
sentence = "Hello Aswathi How are you doing today"
sentence_token = word_tokenize(sentence)
sentence_token
```

```
['Hello', 'Aswathi', 'How', 'are', 'you', 'doing', 'today']
```

3.2. Tipos de tareas: lematización

- **Extirpación de sufijos**, mediante expresiones regulares y diccionarios, para obtener el lema o forma base de una palabra.

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

3.3. Tipos de tareas: etiquetación gramatical

- **Asignación de categoría gramatical** a una palabra e.g. sustantivo, verbo, adjetivo...
- Input para analizador sintáctico
- Clases léxicas: nombres, adjetivos, verbos, adverbios
- Clases funcionales: preposiciones, determinantes, conjunciones, pronombres, interjecciones

3.3. Tipos de tareas: etiquetación gramatical

- En inglés, a una misma palabra se le puede asignar distintas categorías gramaticales en base a su posición y función sintáctica (i.e. ambigüedad gramatical)

  back door (JJ) | on my back (NNS) | to back a proposal (VB) | to come back (RB)

3.3. Tipos de tareas: etiquetación gramatical

- Según la idiosincrasia de cada lengua, se suele seguir un estándar de etiquetado diferente.
- Para el inglés: Penn TreeBank Tagset
- Para el español: EAGLES standard.
- Sistemas basados en reglas o sistemas basados en ML/DL (Hidden Markov Model, Recurrent Neural Network)

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb gerund	*eating*
JJ	adjective	*yellow*	VBN	verb past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	*#*
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Nouns		
nc00000	Unknown common noun (neologism, loanword)	*minidisc, hooligans, re-flotamiento*
nc0n000	Common noun (invariant number)	*hipótesis, campus, golf*
nc0p000	Common noun (plural)	*años, elecciones*
nc0s000	Common noun (singular)	*lista, hotel, partido*
np00000	Proper noun	*Málaga, Parlamento, UFINSA*
Pronouns		
p0000000	Impersonal *se*	*se*
pd000000	Demonstrative pronoun	*éste, eso, aquellas*
pe000000	"Exclamative" pronoun	*qué*
pi000000	Indefinite pronoun	*muchos, uno, tanto, nadie*
pn000000	Numeral pronoun	*dos miles, ambos*
pp000000	Personal pronoun	*ellos, lo, la, nos*
pr000000	Relative pronoun	*que, quien, donde, cuales*
pt000000	Interrogative pronoun	*cómo, cuánto, qué*
px000000	Possessive pronoun	*tuyo, nuestra*
Adverbs		
rg	Adverb (general)	*siempre, más, personalmente*
rn	Adverb (negating)	*no*

3.3. Tipos de tareas: etiquetación gramatical

```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")

for token in doc:
    print(token.text, token.lemma_, token.tag_)
```

RUN

3.3. Tipos de tareas: etiquetación gramatical

```
Apple Apple NNP
is be VBZ
looking look VBG
at at IN
buying buy VBG
U.K. U.K. NNP
startup startup NN
for for IN
$ $ $
1 1 CD
billion billion CD
```

3.3. Tipos de tareas: etiquetación gramatical

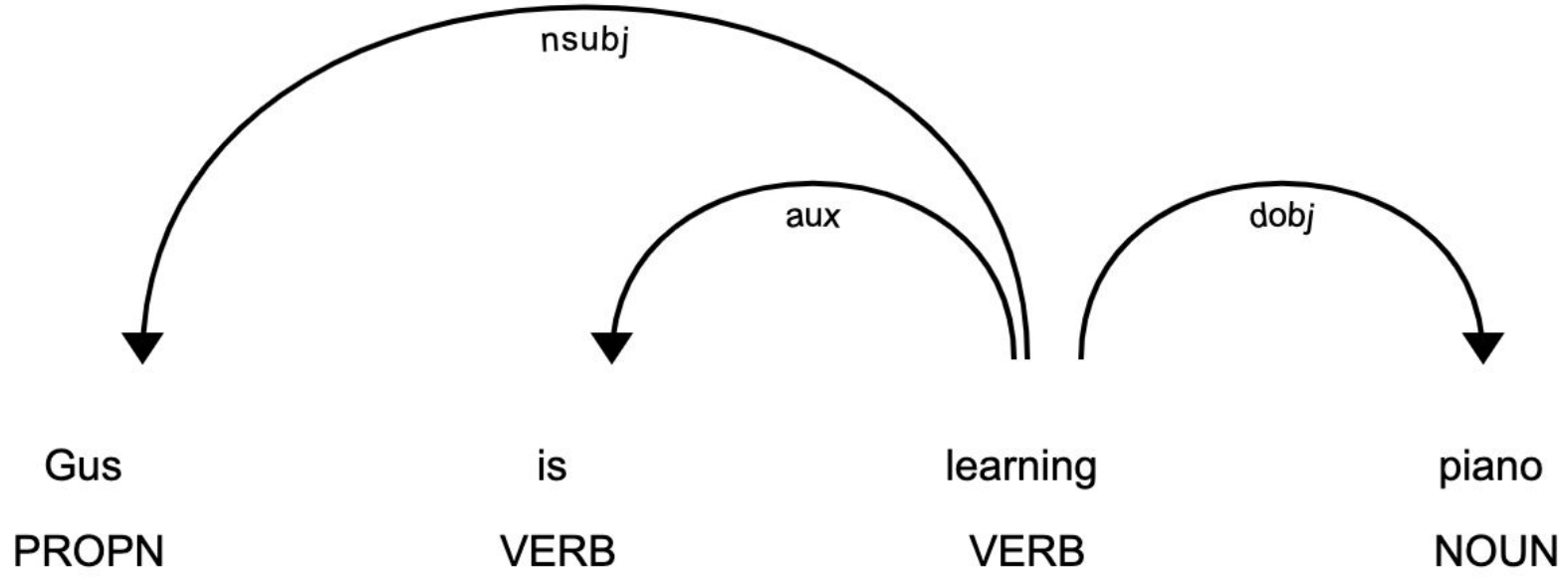
- Características lingüísticas que pueden mejorar el rendimiento (***feature-engineering***):
  - w_i empieza por mayúscula (útil para detectar nombres propios)
  - w_i contiene determinado sufijo (e.g. -ar/-er/-ir en verbos, -al/-able/-ible/-ico, etc.)

3.4. Tipos de tareas: analizador sintáctico

- **Representación**, en ocasiones arbórea, de las relaciones de **dependencia** o **constituencia** entre elementos de una oración.
- Entender la estructura oracional es importante para entender su significado.
  e.g. *Toronto cops kill man with knife* (ambigüedad sintáctica)

- Paso previo en sistemas de extracción de información.
- Sistemas basados en reglas o redes neuronales.

3.4. Tipos de tareas: analizador sintáctico



3.5. Tipos de tareas: NER

- Identificación y **extracción** de **entidades nombradas**: personas, lugares, organizaciones
  e.g. ***Juan*** *sufrió un accidente de coche en* ***Ávila***

- Muy útil en sistemas de extracción de información.
- Sistemas basados en reglas, Machine Learning (Conditional Random Field) o redes neuronales (RNN bidireccional, Transformers).

  ¿Podrías predefinir alguna regla lingüística que ayude a encontrar localizaciones?

3.5. Tipos de tareas: NER

- Se puede valer de otra información lingüística o utilizar una base de datos (***feature-engineering***): tokenización, etiquetación gramatical, base de datos Wikipedia, GeoNames, etc.
- Formato corpus/dataset de entrenamiento y evaluación: datos textuales tabulados, donde cada columna representa una característica, como token, etiqueta gramatical, mayúscula inicial, pertenencia a base de datos, etc. y etiqueta (esquema BIO por ej.)
- Problema: discurso informal en redes sociales

3.5. Tipos de tareas: NER

- LORE: LOcative Reference Extractor (Fernández-Martínez & Periñán, en prensa)
- Sistema para extracción de localizaciones en tweets, basado en expresiones regulares, conocimiento lingüístico (n-gramas, etiquetas gramaticales y reglas), uso de lexicones y diccionarios (GeoNames, EuroWordNet, listas de abreviaturas)
- Multilingüe: métodos semi-automáticos
- F1: 0.81 y 0.75 con dos datasets de evaluación distintos (vs spaCy, NLTK, Stanford NER, Google Natural Language API, OpenNLP)

3.5. Tipos de tareas: NER

- nLORE: neural LOcative Reference Extractor (Fernández-Martínez & Periñán, en preparación)
- Sistema basado en una red neuronal bidireccional RNN con LSTM y CRF para extracción de localizaciones en tweets en inglés
- *Feature-engineering:* token, etiquetas gramaticales, GeoNames, WordNet, marcador locativo
- Corpus de entrenamiento (7000 tweets), corpus de validación (1050 tweets), corpus de evaluación (1350 tweets)
- F1: 0.79 vs 0.75 LORE
- Contribución características lingüísticas extra: mejora muy ligeramente (0.05)

3.5. Tipos de tareas: NER

Flooding	NN	O
across	IN	O
parts	NNS	O
of	IN	O
Broward	NNP	B_LOCATION
County	NNP	E_LOCATION
to	TO	O
improve	VB	O
throughout	IN	O
day	NN	O

3.6. Análisis de sentimiento

- También llamado minería de opinión.
- **Análisis** de la **polaridad** de una opinión acerca de un producto, una marca, una persona, una idea, etc.

  e.g. *El móvil me **funciona a la perfección***
  e.g. *La batería me **dura poquísimo***
  *e.g. Las políticas de X nos están llevando a la **ruina***
  e.g. *Me parece una película **estupenda**, pero el argumento **no está a la altura***

3.6. Análisis de sentimiento

- Enfoques basados en reglas y lexicones:
  - Listas de palabras positivas y negativas
  - Contar número de ocurrencias en texto/oración, y las que prevalezcan determinan la polaridad
  - Enfoque pobre en cuanto a semántica y pragmática: ironía, sarcasmo

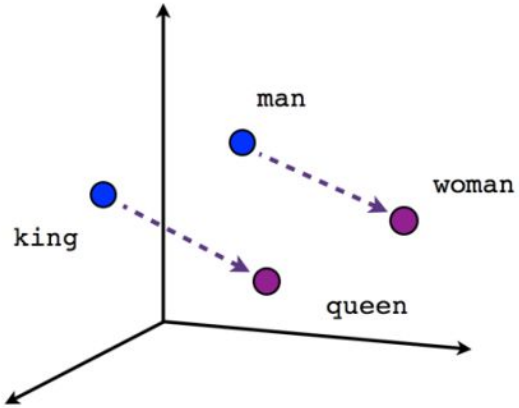
- Enfoques basados en ML:
  - *Bag of words* o extracción/vectorización de características textuales
  - Algoritmo de clasificación: SVM, Naïve Bayes, Regresión lineal
  - Redes neuronales: CNN

3.7. Word embeddings (similitud semántica)

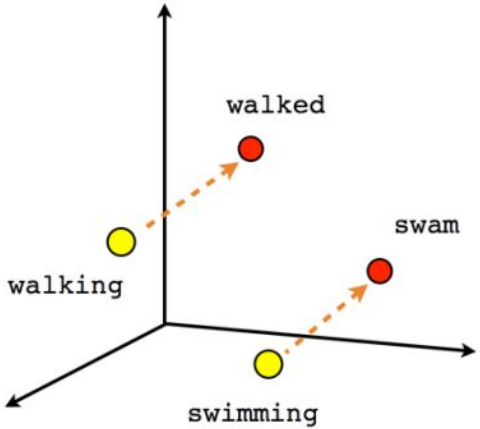
- ¿Cómo definir el **significado** de una palabra?
  “You shall know a word by the company it keeps” (Firth, 1957)

- Word embeddings: **vectorización** de palabras o frases para cuantificar o categorizar la **similitud semántica** entre estas en un espacio vectorial mediante el cálculo de la distancia del coseno (Mikolov et al., 2013)
- Capturan información sintáctica y semántica.
- Librerías y herramientas: Word2vec, GloVe
- Para entrenamiento, corpus en crudo sin etiquetar.

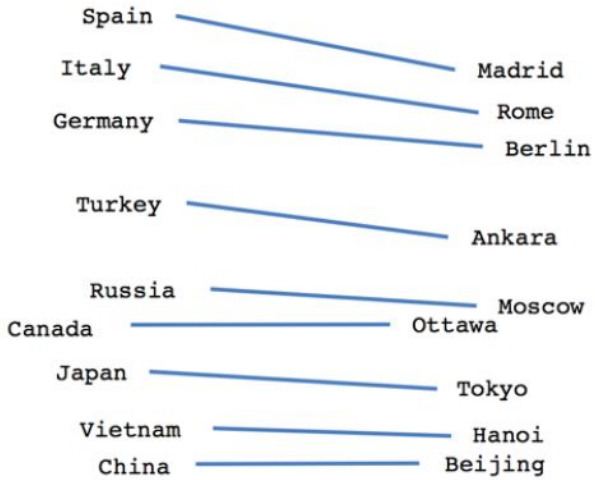
3.7. Word embeddings (similitud semántica)



Male-Female

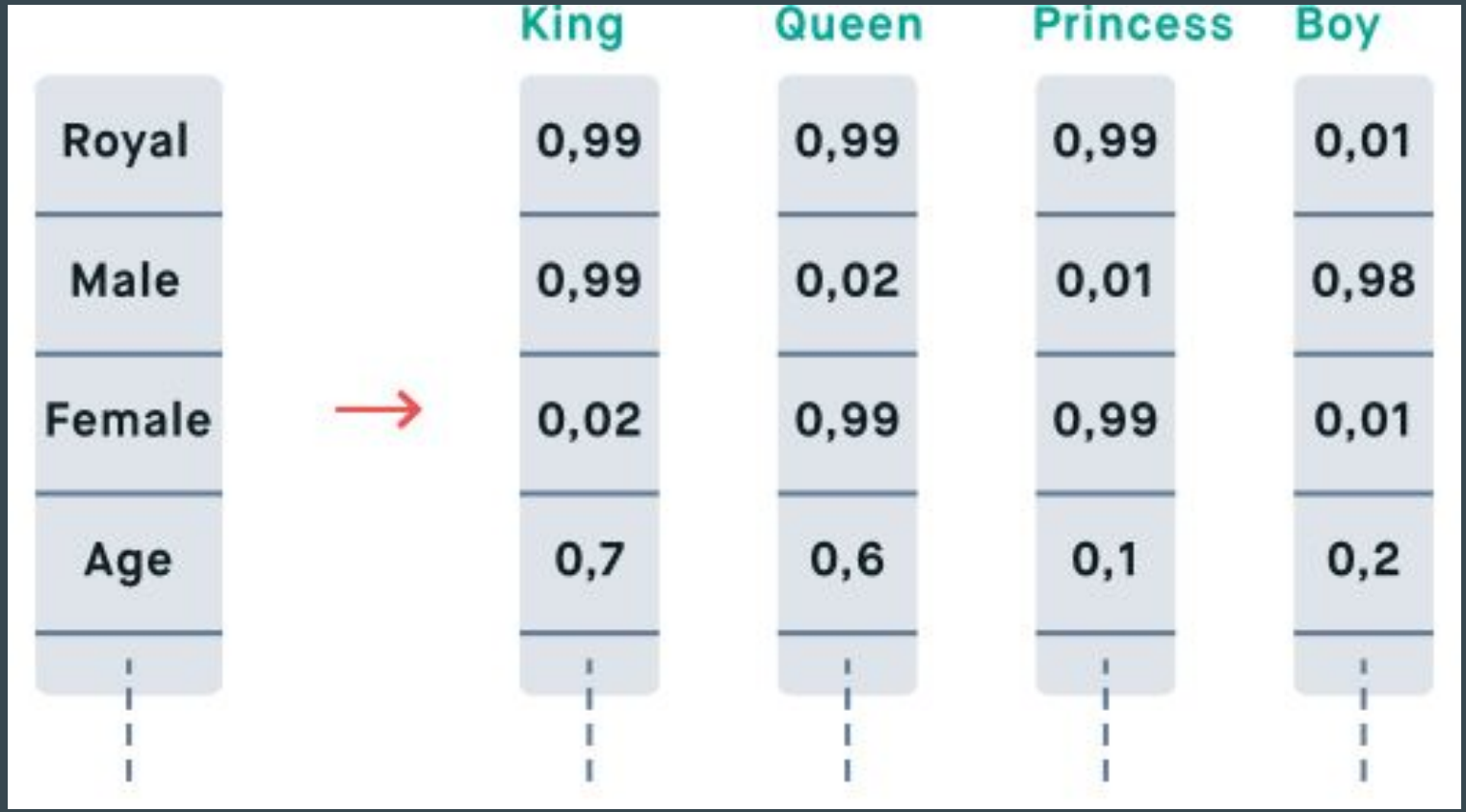


Verb tense



Country-Capital

3.7. Word embeddings (similitud semántica)



3.8. Tipos de tareas: detección del tema

- **Descubrir** el **tema** recurrente o más frecuente de una oración, párrafo, texto o colección de textos en base a una lista predefinida de temas, o agrupar textos en base a su similitud temática.
  e.g. Cae la cuota de mercado de la compañía X ⟶ economía
- Se parte de palabras clave, o sin información previa: esto determina tipo de algoritmo
- Sin información previa:
- Sistemas basados en ML: Latent Semantic Analysis (bag of words, frecuencias, tf-idf)

3.8. Tipos de tareas: detección del tema

- Partiendo de unos temas predefinidos:
- Input:
  un corpus de documentos d
  un conjunto de clases $C = \{c_1, c_2, ..., c_i\}$
- Output: una clase predicha $c \in C$

3.8. Tipos de tareas: detección del tema

- Sistemas basados en reglas y lexicones (frecuencias de palabras pertenecientes a clases léxicas, campos léxicos y relaciones de hiponima, meronimia, sinonimia, etc.)
  e.g. un texto sobre economía incluye términos relacionados al ámbito de la economía (el género puede ser un artículo periodístico, un artículo científico, etc.): *comprar, vender, mercado, bienes, servicios, cliente, técnicas de mercadotecnia, atraer inversión extranjera*, etc.
- Sistemas basados en ML: Naive Bayes (bag-of-words como característica), SVM, redes neuronales enriquecidas con word embeddings(RNN o CNN)

3.9. Tipos de tareas: resumidor automático

- **Proporcionar** una **síntesis** de las **ideas** más importantes contenidas en un texto o colección de textos
- Convergencia entre NLP y NLG
- Resumidores extractores (palabras y expresiones tomadas de estos textos) y resumidores abstractos (paráfrasis y reformulación)
- Resumidor extractores, sistemas basados en...
  - Reglas y conocimiento lingüístico:
    e.g. Ana, l~~a poeta que por entonces vivía en Madrid~~, se inspiró en el realismo sucio de Bukowski (quitar NP aposicionales)
  - Redes neuronales (RNNs)

3.9. Tipos de tareas: traducción automática

- Convertir/**transformar** un texto fuente en un idioma X a un texto meta en el idioma Y. NLP y NLG.
- No se trata de traducir palabra por palabra: **retención** del **significado**.
- **Interdependencia** de todos los **niveles lingüísticos**, destacando la sintaxis, la semántica y la pragmática, y conocimientos extralingüísticos (conocimiento sociocultural) en ambas lenguas.
- Traductores automáticos basados en...
  - diccionarios, reglas y lexicones, una interlengua semántica: muy costoso y complejo, pero eficiente
  - redes neuronales (Neural Machine Translation): RNN, attention-based mechanisms

3.9. Tipos de tareas: traducción automática

El niño estiró la pata	The boy kicked the bucket
El niño estiró la pata para golpear el balón	The boy stretched out his leg to hit the ball
El niño estiró la pata cuando le golpeó una piedra	The kid kicked the bucket when he got hit by a rock
La situación no hay por dónde cogerla	There's nowhere to take the situation

3.10. Agentes conversacionales (chatbots)

- Asistentes virtuales que permiten **comunicación**, **interacción** o **conversación** entre humano y máquina: toman una pregunta/orden/afirmación como input y devuelven una respuesta como output.
- **Convergencia** de distintas **disciplinas** (NLP, NLG) y convergencia de distintas **tareas** (etiquetación gramatical, análisis sintáctico, NER, extracción de información, minería de opinión, detección del tema...)

3.10. Agentes conversacionales (chatbots)

- **Proceso complejo**:
- Entender **input** (speech to text): reconocimiento de voz, sintaxis, semántica léxica (significado literal) y pragmática (intención del hablante)
- Determinar tipo de **output**: una afirmación, una pregunta, consejo, etc.
- Producir **output** (text to speech): respuesta fluida y natural

3.10. Agentes conversacionales (chatbots)

- Tipos:
  - Basados en reglas y conocimiento lingüístico: simples, pero costoso a nivel de conocimientos y complejidad lingüísticos
  - Basados en ML (RNN, LSTM, CRF...): más sofisticados, pero cantidad inmensa de datos de entrenamiento
  - Híbridos: lo mejor de ambos
- Limitaciones:
  - Etiquetado masivo y complejo para dataset de entrenamiento
  - Pragmática: Capacidad pobre para discernir el significado de acuerdo al contexto e.g. Hace mucho frío (orden indirecta para activar la calefacción)

3.10. Agentes conversacionales (chatbots)

- Ejemplos: Siri, Google Assistant, Cortana...
  https://dialogflow.com/

4. Limitaciones

- Escasez de datasets etiquetados lo suficientemente grandes y representativos.
- La evaluación de los sistemas de NLP se lleva a cabo en condiciones poco rigurosas.
- Representación del conocimiento semántico enciclopédico (vs conocimiento semántica distribucional): e.g. *robar* (concepto STEAL), *ir al supermercado*
- Lenguaje informal en redes sociales

4. Limitaciones

- Contexto extralingüístico (cultura, sociedad): e.g. traducción automática
- Pragmática (intención del hablante): ironía, sarcasmo, órdenes indirectas, etc.

5. Conclusiones

- NLP es capaz de resolver con gran precisión muchas tareas relacionadas con la comprensión del lenguaje humano.
- Todavía, sin embargo, se encuentra muy lejos de comprender la naturaleza del lenguaje humano.
- Reto: conseguir (e incluso superar) las habilidades lingüísticas y cognitivas del ser humano, y simular la naturaleza del lenguaje humano

6. Bibliografía

- Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*, (May), 48–57.
- Chiticariu, L., Li, Y., & Reiss, F. R. (2013). Rule-based information extraction is dead! Long live rule-based information extraction systems! In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 827–832).
- Fernández-Martínez, N.J. & Periñán-Pascual, C. (unpublished). A linguistically-aware model for microtext location detection.
- Fernández-Martínez, N.J. & Periñán-Pascual, C. (in preparation). nLORE: A linguistically-rich neural implementation of LOcative Reference Extractor..
- Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing*. Unpublished.
- Periñán-Pascual, C. (2012). En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica. *Onomázein*, *26*, 13–48. Retrieved from http://www.fungramkb.com/resources/papers/022.pdf
- Riemer, N. (2010). *Introducing Semantics*. New York: Cambridge University Press.