Reconocimiento de Entidades Nombradas en Mensajes Cortos

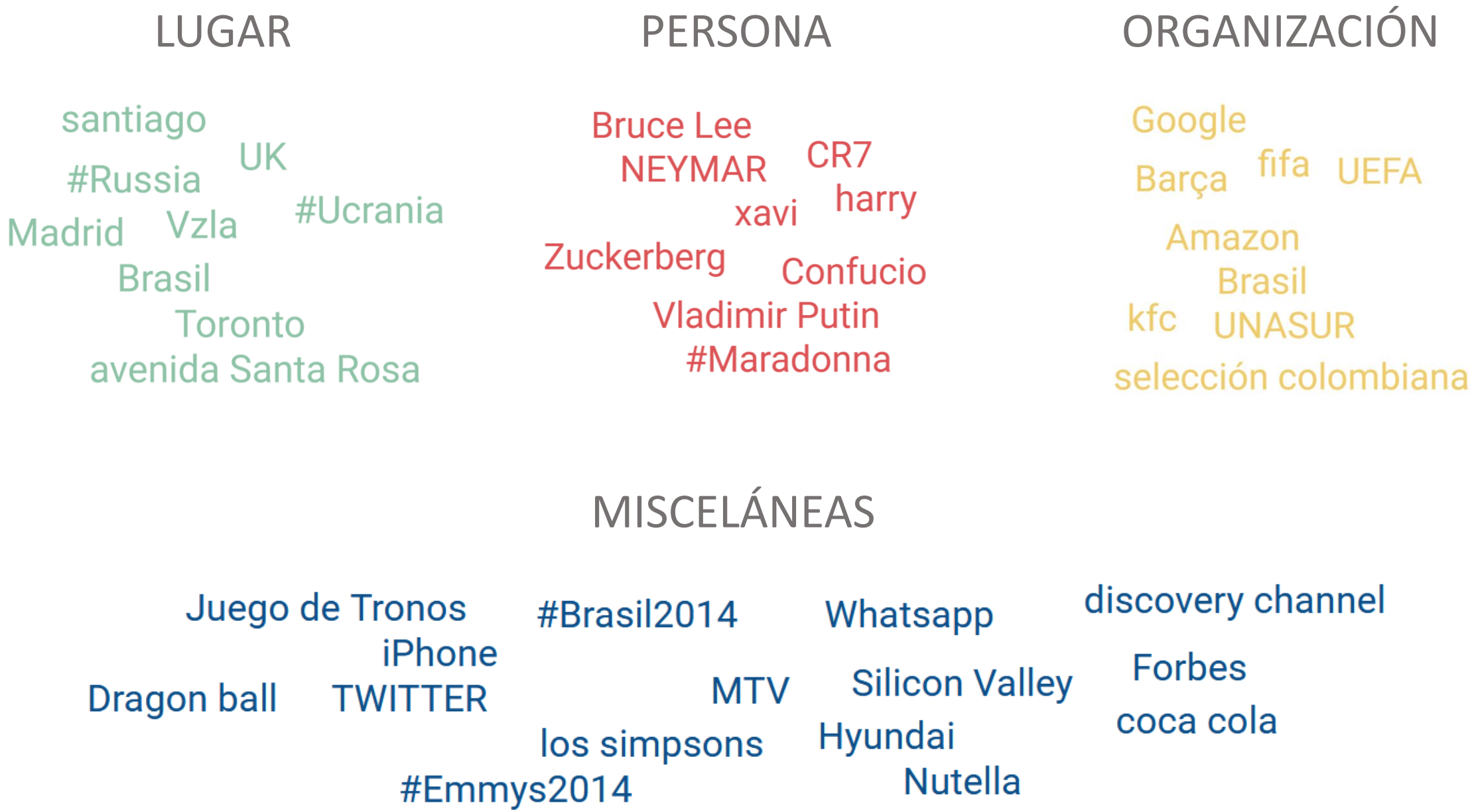
Autor: **Laila González Fernández**

Tutores: **Dr. Yudivián Almeida Cruz**

MSc. Suilán Estévez Velarde

Facultad de Matemática y Computación
Universidad de La Habana

- Inmediatez
- Gran número de mensajes
- Variedad


LUGAR
santiago
UK
#Russia
#Ucrania
Madrid
Vzla
Brasil
Toronto
avenida Santa Rosa
PERSONA
Bruce Lee
NEYMAR
CR7
xavi
harry
Zuckerberg
Confucio
Vladimir Putin
#Maradonna
ORGANIZACIÓN
Google
Barça
fifa
UEFA
Amazon
Brasil
kfc
UNASUR
selección colombiana
MISCELÁNEAS
Juego de Tronos
#Brasil2014
Whatsapp
discovery channel
iPhone
Forbes
Dragon ball
TWITTER
MTV
Silicon Valley
coca cola
los simpsons
Hyundai
#Emmys2014
Nutella

OnCuba @OnCuba · 5d

Danza Contemporánea de #Cuba fascina en #ReinoUnido. oncubamagazine.com/cultura/danza-...

Danza Contemporánea de #Cuba fascina en #ReinoUnido.

ORGANIZACIÓN **LUGAR**

Santiago de Cuba Holguín

Brasil (país) Brasil (equipo de fútbol)
Lugar Organización

xavi Vladimir Putin TWITTER

#Brasil2014 @Cristiano

Venzuela Vzla

VS

Proponer y diseñar una metodología para el desarrollo de un sistema capaz de reconocer entidades nombradas de diversos dominios en mensajes cortos en español.

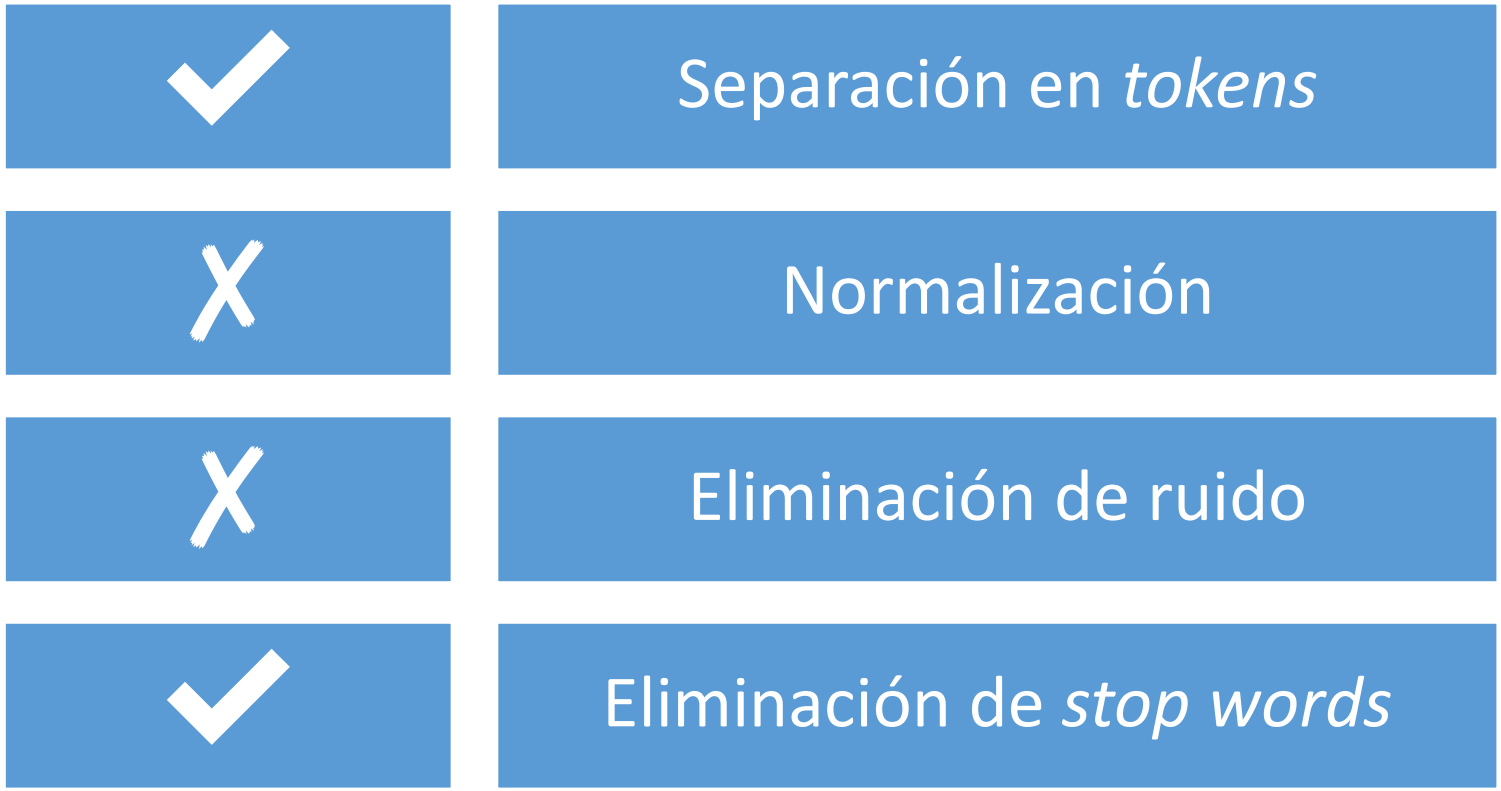
- Hacer un estudio del estado del arte concerniente al Reconocimiento de Entidades Nombradas.
- Identificar las características del Reconocimiento de Entidades Nombradas en mensajes cortos.
- Desarrollar una metodología que no requiera grandes cantidades de datos anotados.
- Desarrollar una metodología capaz de emplear bases de conocimiento distintas según el dominio.
- Evaluar la eficacia de la metodología desarrollada en Twitter y otros dominios.

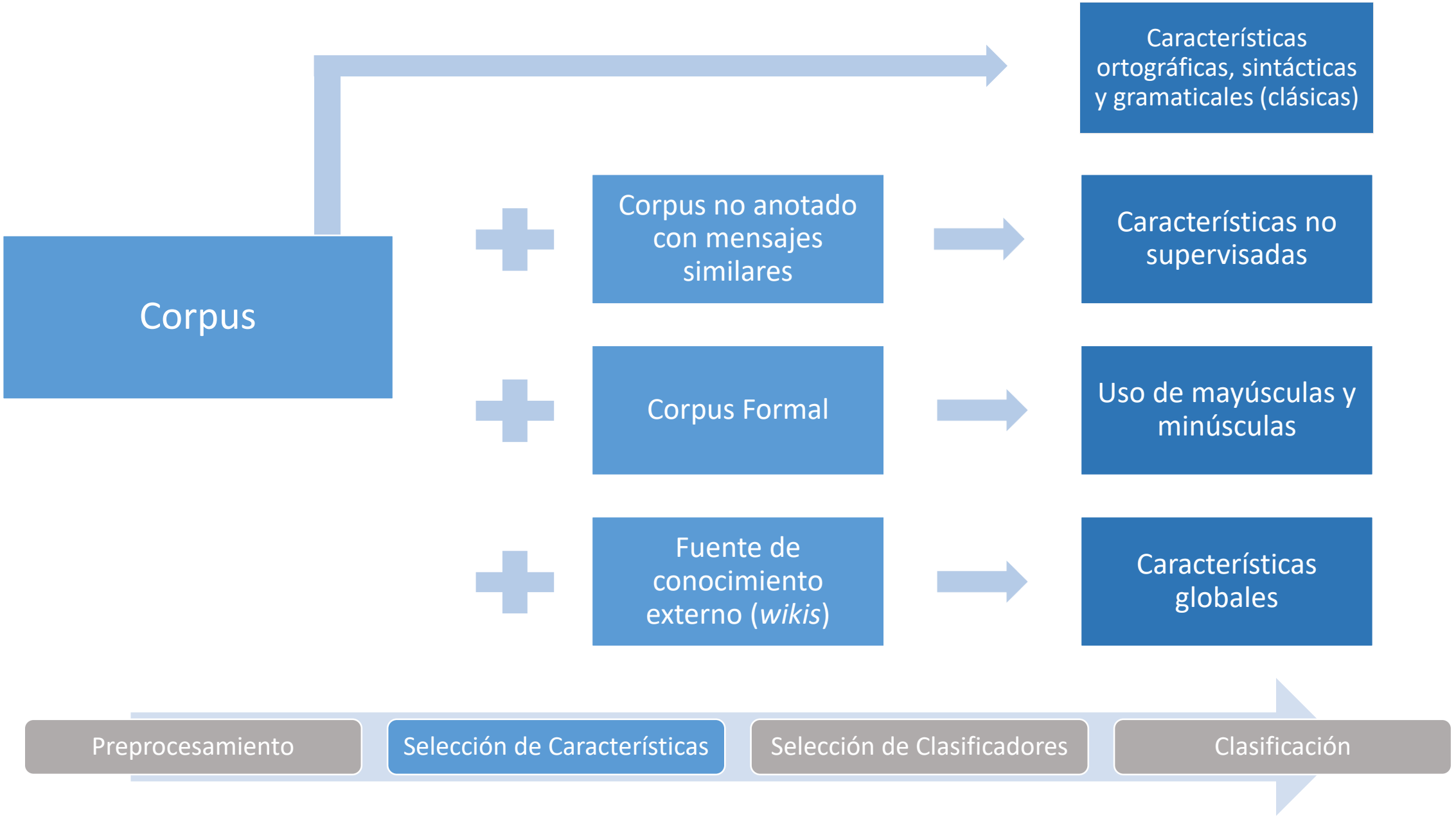

✗
Reglas
✗
Bases de conocimiento
Aprendizaje de Máquinas
✗
Supervisado
✗
No Supervisado
✓
Semi Supervisado

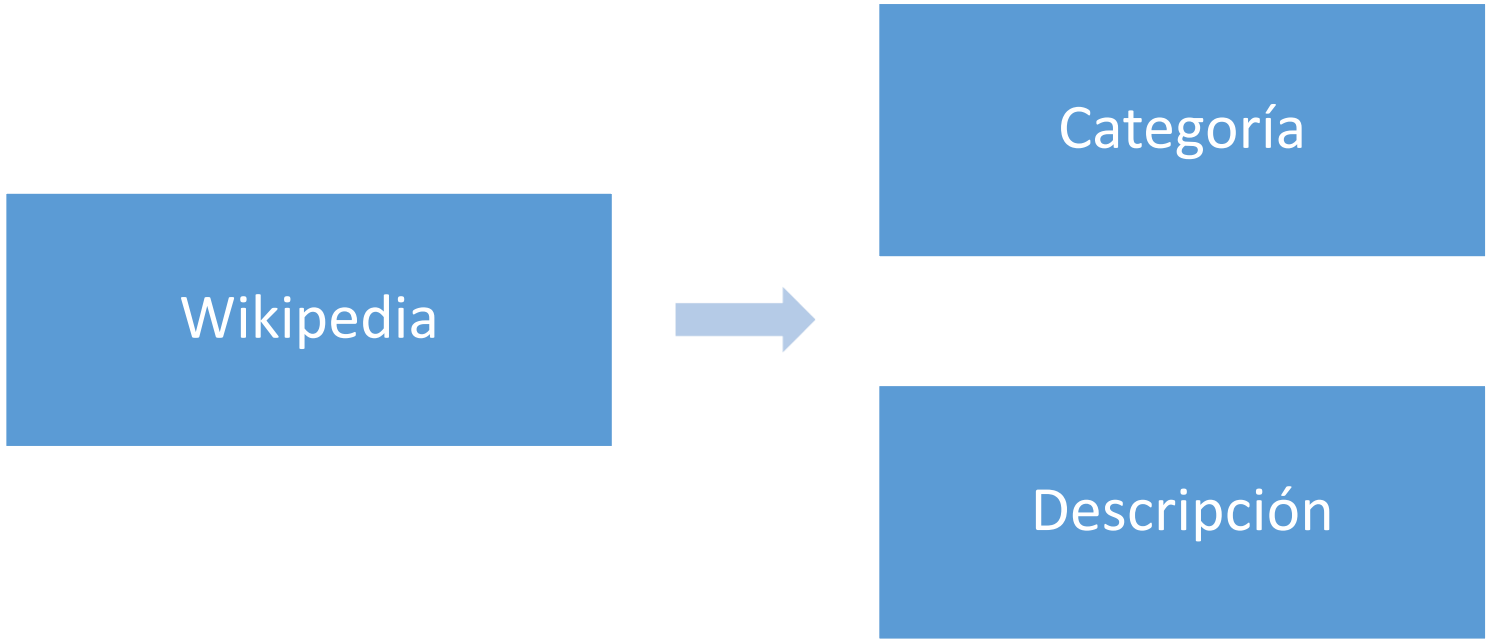
Self-Training

Co-Training


Preprocesamiento
Selección de Características
Selección de Clasificadores
Clasificación


Separación en *tokens*
Normalización
Eliminación de ruido
Eliminación de *stop words*


Corpus
Corpus no anotado con mensajes similares
Corpus Formal
Fuente de conocimiento externo (*wikis*)
Características ortográficas, sintácticas y gramaticales (clásicas)
Características no supervisadas
Uso de mayúsculas y minúsculas
Características globales
Preprocesamiento
Selección de Características
Selección de Clasificadores
Clasificación



Rusia (en ruso: Россия) es el **país** más extenso del mundo.

Clasificadores tradicionales

Clasificadores para datos estructurados

- Cadenas Ocultas de Markov (HMM)
- Modelos de Máxima Entropía (Maximum Entropy)
- Campos Aleatorios Condicionales (CRF)

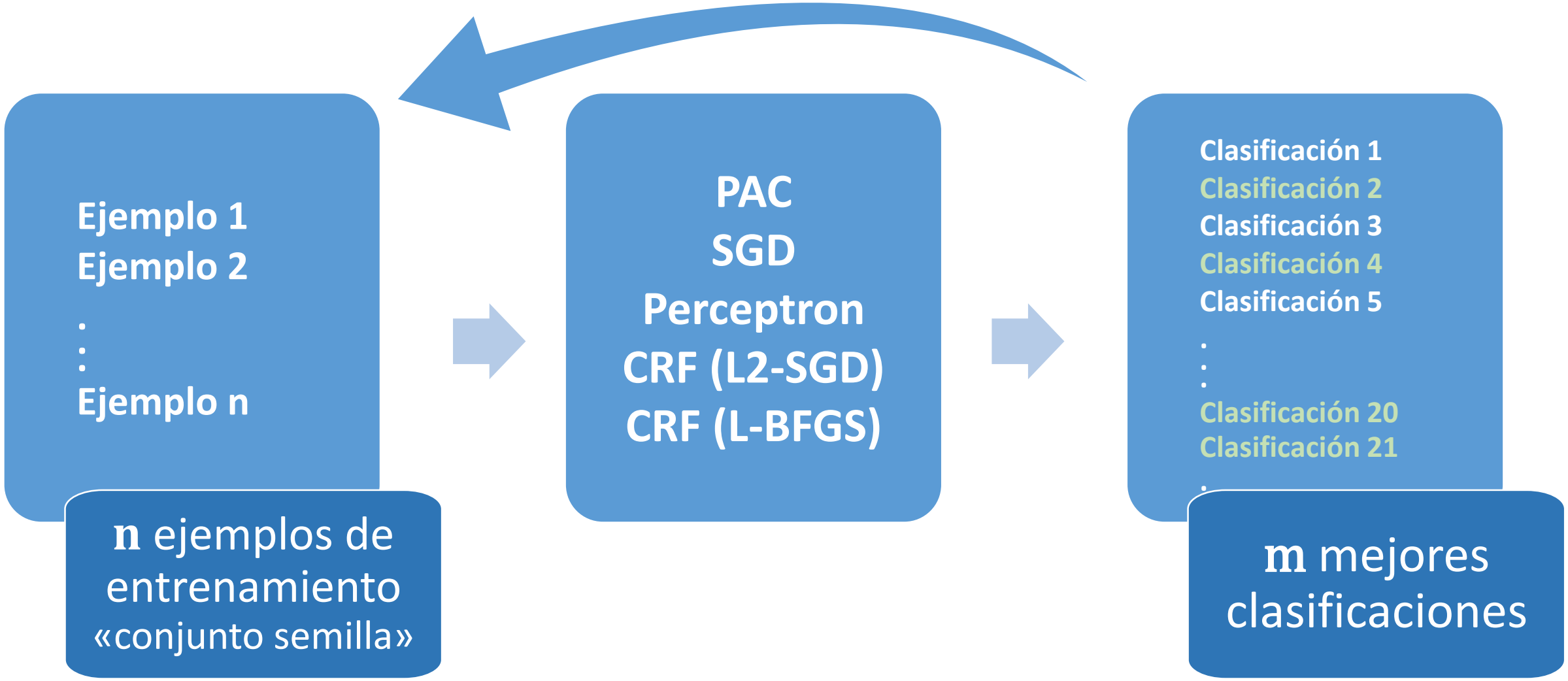
Clasificador	Precisión	Exhaustividad	Medida F1
Línea de base	0.006	0.006	0.006
Naive Bayes	0.552	0.266	0.324
Árboles de Decisión	0.320	0.401	0.352
SVM	0.283	**0.553**	0.370
SGD	**0.589**	0.412	0.457
Perceptron	0.481	0.435	0.436
PAC	0.533	0.426	**0.458**
AdaBoost	0.479	0.308	0.348
Random Forest	0.564	0.303	0.383

SVM: Support Vector Machine, SGD: Stochastic Gradient Descent, PAC: Passive Agressive Classifier

Tipo de Entidad	Precisión	Exhaustividad	Medida F1
LUGAR	0.541	0.464	0.464
PERSONA	**0.625**	**0.555**	**0.579**
ORGANIZACIÓN	0.449	0.270	0.327
MISCELÁNEA	0.334	0.141	0.189
NO ENTIDAD	0.986	0.995	0.990

Clasificador	Precisión	Exhaustividad	Medida F1
CRF (L-BFGS)	0.626	0.475	**0.517**
CRF (L2-SGD)	**0.635**	0.459	0.509
CRF (AP)	0.548	**0.465**	0.490
CRF (PA)	0.557	0.547	0.488
CRF (AROW)	0.481	0.412	0.434
Maximum Entropy	0.520	0.381	0.427
HMM	0.054	0.136	0.077

Self-Training



n	m	Precisión	Exhaustividad	Medida F1
500	500	0.516	0.365	0.370
500	1000	0.562	0.338	0.389
1000	500	0.540	0.410	0.438
1000	1000	0.602	0.365	0.428
2000	500	0.585	0.453	**0.495**
2000	1000	0.600	0.408	0.468
3000	500	0.509	**0.476**	0.482
3000	1000	**0.617**	0.433	0.494

n: tamaño del conjunto de entrenamiento inicial

m: número de mensajes a añadir al conjunto de entrenamiento en cada iteración

Sistema	Precisión	Exhaustividad	Medida F1
Stanford NER	0.299	**0.456**	0.361
Stanford NER (entrenado)	**0.627**	0.292	0.398
T-NER	0.319	0.281	0.293
Propuesta	0.602	0.408	**0.438**

Mejoría:

10% respecto al *Stanford NER Tagger* entrenado

21% respecto al *Stanford NER Tagger* para el idioma español

49% respecto a *T-NER*

Tipo de Entidad	Precisión	Exhaustividad	Medida F1
LUGAR	0.735	0.634	0.675
PERSONA	**0.816**	**0.858**	**0.834**
ORGANIZACIÓN	0.710	0.799	0.751
MISCELÁNEA	0.597	0.494	0.499
NO ENTIDAD	0.991	0.984	0.987

Corpus CoNLL 2003
Medida F1: 0.723

Sistema	Precisión	Exhaustividad	Medida F1
CMP	0.813	0.814	0.814
Flo	0.780	0.794	0.791
CY	0.781	0.761	0.772
Stanford NER	0.780	0.762	0.771
WNC	0.759	0.774	0.766
BHM	0.742	0.774	0.758
Tjo	0.760	0.756	0.758
PWM	0.743	0.735	0.739
Jan	0.740	0.738	0.739
Mal	0.739	0.734	0.737
Propuesta	0.718	0.729	0.723
Tsu	0.690	0.741	0.715
BV	0.605	0.673	0.637
MM	0.563	0.665	0.610
Línea de base	0.263	0.565	0.359

Conclusiones

- Se propone una metodología utilizando *self-training* para el reconocimiento de entidades en mensajes cortos.
- Se observan mejorías respecto a otras metodologías existentes .
- La propuesta necesita solo un pequeño número de mensajes anotados para su entrenamiento y es una propuesta portable a otros dominios.

Recomendaciones

- Estudiar la elección de clasificadores para el proceso de *self-training* y de los valores adecuados para sus parámetros.
- Utilizar otras técnicas de aprendizaje semi-supervisado.
- Aplicar la metodología en otros dominios.
- Evaluar el impacto de enriquecer el corpus con oraciones tomadas de textos estructurados.

Reconocimiento de Entidades Nombradas en Mensajes Cortos

Autor: **Laila González Fernández**

Tutores: **Dr. Yudivián Almeida Cruz**

MSc. Suilán Estévez Valverde

Facultad de Matemática y Computación
Universidad de La Habana

¿Las características que en su propuesta son consideradas como las más acertadas en la predicción de cada clase se mantienen al cambiar el dominio?

Características	Precisión	Exhaustividad	Medida F1	%
Clásicas	0.502	0.358	0.418	95%
No supervisadas	0.250	0.155	0.162	37%
Globales	0.462	0.270	0.319	73%
Todas	0.540	0.410	0.438	100%

Corpus de Tweets

Características	Precisión	Exhaustividad	Medida F1	%
Clásicas	0.697	0.652	0.694	96%
No supervisadas	0.423	0.228	0.289	40%
Globales	0.539	0.391	0.448	62%
Todas	0.718	0.729	0.723	100

Corpus de Noticias

Clase	Característica
No entidad	t.first
No entidad	t.shape: -
No entidad	t.shape: aa
Lugar	wiki_category: Estados miembros de la ONU
Persona	wiki_category: Nombres
Lugar	wiki_description: ciudad
No entidad	t.postag: pronombre
Lugar	t-1: en
Persona	wiki_category: Personas
Persona	wiki_description: futbolista

Corpus de Tweets

Clase	Característica
No entidad	t.shape: aa
No entidad	t.shape: -
Lugar	wiki_description: capital
Lugar	t-1: en
Lugar	wiki_description: ciudad
Persona	wiki_category: Personas
Persona	wiki_category: Apellidos
No entidad	t.first
Lugar	wiki_description: pueblo
Miscelánea	t.shape: AA00

Corpus de Noticias

Una de las ventajas de los algoritmos de aprendizaje de máquina es la posibilidad de clasificar entidades nombradas no presentes en bases de conocimiento, ni clasificadas previamente en la fase de entrenamiento. ¿Qué resultados serían obtenidos por su propuesta en estos casos? ¿En estos casos cuál es la característica que mejor los identifica?

Entidades	**Precisión**	**Exhaust.**	**Medida F1**	**%**
No vistas en entrenamiento	0.453	0.323	0.349	80%
No encontradas en bases de conocimiento	0.531	0.363	0.410	94%
No encontradas en bases de conocimiento ni vistas en entrenamiento	0.414	0.240	0.278	63%
Todas	0.540	0.410	0.438	100%

GameCaptureHD
Ultraport
Johaaaaan
Diomedez
SportLeon
#Casselton
Flappy
#Lanus
CALASANZ
VITORIA
Protón-M
#BEYONC

¿Cómo se incorpora en el sistema el aprendizaje no supervisado?

Algoritmos de *clustering*

- *Clusters* de Brown
- *Clusters* de Clark
- Word2Vec


YoestoyconChavez
@lourdeszuazo
@Mariaeffarias
@odinmiranda
@maverik
Michael
Maduro
maduro
erPlano
Lula
Noruega
URGENTE
Cristina
Cabello
Segunda
Sociales
Chaderton
@oswaldopaya
Ahmadineyad
Putin
Ahmadinejad
@MariVelBa


Bolivia
Twitter
Guantánamo
México
España
Colombia
Chile
China
Latinoamérica
Brasil
Cuba
Marzo
Siria
F
Ecuador
Camagüey
cuba
Rusia
abril
EEUU
Nicaragua
Venezuela
Miami
ONU
América
Madrid
Bolívar
Europa
Argentina
Cienfuegos


Oposición
@CFKArgentina
CELAC
Celac
ONU
Instituciones
@CNNEE
UNEAC
Embajada
Copa
#MCL
Canciller
#AsambleaNacional
pidió
India


reconocimiento
clasifica
vence
expulsa
especialmente
entrará
arrolló
Proponen
fieles
vienes
rumbo
llegan
ofensa
volverse
despedida
acompañó
alienta
vino
busca
respaldo
derrotó
pone
llegado
van
despide
propone
invito
llegó
llevó
empezar
enviar
volver
regresar
viajará
vuelve
dedicada
felicita
respaldar
excluir
vas
regresa
link
honran
retando
noqueó
llama
insulte
boleto
elimina
niega
vamos
llamado
abrazar
despiden
responde
exhorta
volvió
vuelven
pronuncia
vengas
acceso
entrar
convoca
llegará
explicarse
homenajes
parió
disfrutar
llegué
pese
fúnebres
antiterroristas
#Guadalupe
Protección

Clark:

Cluster 12: **Cuba**, Bolivia, cuba, islajuventud, Egipto, SanctiSpíritus, VenezuelaDecide

Word2vec:

Cluster 58: **Cuba**, Caribe, Cubarte, bahía, israel, Cienfueguero

Brown

Cluster 01010001100 : **Cuba**, España, China, América, México, Venezuela, Bolívar
Cluster 010100011: **Cuba**, …, Guatemala, Washington, #España, Caracas
Cluster 0101000: **Cuba**, …, Naciones, Estado, Egipto, #habana, Villa, estos

clark: 12, word2vec: 58, brown-11: 01010001100,
brown-9: 010100011 brown-7: 0101000

Características	Precisión	Exhaustividad	Medida F1	%
Clásicas	0.532	0.391	0.436	95%
No supervisadas	0.250	0.155	0.169	37%
Globales	0.462	0.270	0.329	72%
Todas	0.533	0.426	0.458	-

Corpus de Tweets

Reconocimiento de Entidades Nombradas en Mensajes Cortos

Autor: **Laila González Fernández**

Tutores: **Dr. Yudivián Almeida Cruz**

MSc. Suilán Estévez Valverde

Facultad de Matemática y Computación
Universidad de La Habana