Compendio del foro de evaluación IberLEF 2019

Informe técnico TR 2020-1/5511297[1]

Ana García Serrano (agarcia@lsi.uned.es)

ETSI Informática - UNED

Contenido

Febrero 2020

[1] Citar: Garcia Serrano, A. "Compendio sobre el foro de evaluación en español IberLEF2019" Informe Técnico TR 2020-1/5511297. Disponible en e-spacio UNED (http://e-spacio.uned.es/fez/).

1. Introducción

Este resumen que intenta ser conciso y sustancial, es consecuencia de las aportaciones realizadas en una actividad de aprendizaje, en el marco de la asignatura "Semántica y pragmática en la web" del máster de Tecnologías de la Lengua de la UNED[2], en el curso 2019-20, a partir del contenido disponible *on-line*[3] del foro *Iberian Languages Evaluation Forum* IberLEF2019 (organizado en el marco de la SEPLN2019).

IberLEF2019 es un foro de evaluación en el que se plantean retos o tareas competitivas de procesamiento de textos para las lenguas de la península ibérica (español, portugués, catalán, vasco y gallego). Este foro de evaluación está organizado a modo de competición entre los sistemas participantes que asumen un mismo reto, esto es la realización de una tarea o resolución de un problema con los mismos datos y en el mismo escenario. Los organizadores del reto deben aportar un *dataset* o corpus, definir el reto o tarea a resolver, indicar las medidas de evaluación de los resultados para que estos puedan ser comparados y en ocasiones se encargan del repositorio de reproducibilidad de los sistemas participantes con las diferentes aproximaciones.

IberLEF 2019 ha consistido en las nueve líneas de trabajo o retos siguientes: descubrimiento de conocimiento en salud (eHealth-KD), análisis de eventualidad y clasificación (FACT), análisis de humor basado en anotaciones humanas (HAHA), detección de ironía en variantes del castellano (IroSvA), negación en castellano (NEGES), reconocimiento de entidades nombradas y extracción de relaciones en portugués (NER Portugués), detección de autoría y agresividad en Twitter para la variante del español en México (MEX-A3T), análisis de sentimiento (TASS) y anonimización de documentos médicos (MEDDOCAN).

Las aproximaciones utilizadas en los sistemas participantes de la mayoría de las tareas o retos se basan en el conocimiento (lingüístico, ontológico u otros) o en el aprendizaje automático. En este segundo caso, además de aproximaciones de aprendizaje automático tradicional como son SVM o *Random Forest*, hay técnicas de aprendizaje profundo, como BERT o ELMO (redes de neuronas). Además, hay un tercer tipo de aproximaciones de modelos del lenguaje basados en *embeddings* (vectores que abren un poco la caja negra de las redes de neuronas), como son *FastText, Glove o Wikipedia2vec*.

El aprendizaje profundo (muy básicamente) consiste en entrenar una red de neuronas para clasificar y ordenar soluciones [51]. El conjunto de datos de entrenamiento puede estár anotado y la red está organizada en torno a una arquitectura. Las arquitecturas más referenciadas entre los participantes de IberLEf2019 son CRF, CNN (que aprende por capas), LSTM y BiLSTM. En la arquitectura *Transformer* (desarrollada en el lenguaje *Python*[4] y liberada por *Google*), el entrenamiento es exageradamente costoso, por eso los investigadores la utilizan mediante sus modelos pre-entrenados como BERT (que se libera en 2018), un modelo del lenguaje con *Word Embeddings* contextuales pre-entrenados. El lector observará en el

[2] http://portal.uned.es/portal/page?_pageid=93,69878406&_dad=portal&_schema=PORTAL&idAsignatura=3110131-&idTitulacion=310701

[3] http://ceur-ws.org/Vol-2421/

[4] https://es.wikipedia.org/wiki/Python

siguiente apartado que en todos los retos algún participante, si no lo son todos, utiliza modelos de aprendizaje profundo de forma aislada o combinada.

Como este texto tiene un objetivo fundamentalmente didáctico, a toda la tecnología nombrada se le asocia bibliografía o links de páginas con más información, la primera vez que aparece (a partir de esta introducción).

Este tipo de tareas compartidas en foros de evaluación son muy importantes en la investigación, a falta de plataformas de evaluación desarrolladas para tareas concretas [30]. Su organización es una estrategia utilizada desde hace tiempo en el área de la recuperación de información[5] o extracción de información[6] con colecciones documentales o multimedia digitales. Plantean problemas abiertos y sus recursos para la investigación son compartidos por los participantes, lo que permite la discusión y un avance en el estado del arte.

Una de las primeras iniciativas en esta área es *Message Understanding Conference (MUC)*, que se organizaba en torno a documentos que trataban sobre operaciones navales (1987 y 1989), terrorismo en países latinoamericanos (1991) o informes de lanzamiento de satélites (1998). Otra competición con mucho prestigio es SemEval[7] para la evaluación de sistemas de análisis semántico[8] y pragmático, que comenzó en 2010 a partir de Senseval[9], que desde 1997 se centraba en la evaluación de sistemas de desambiguación de la semántica de las palabras (*Word Sense Disambiguation*, WSD[10]). Entre las iniciativas europeas se encuentran CLEF (*Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum*)[11], que comienza en el 2000 y en el 2010 se redefine a su estado actual [15] o el foro MediaEVAL[12] (*The Multimedia Evaluation Benchmark)*, que también comienza en 2010, tras dos ediciones de VideoCLEF (2008 y 2009).

En estas tareas compartidas es muy importante el papel de los organizadores porque deben generar interés para tener participantes lo que consiguen, por ejemplo, planteando problemas difíciles, hipótesis de investigación no comprobadas (como en IberLEF2019 hicieron en FACT, IroSva o TASS), modificando los enunciados de las tareas a lo largo de los años o en la misma edición como se hizo en NER_Portuguese, o bien facilitando la participación con atención eficaz durante el periodo de realización de la tarea (habitualmente acotado y con plazos estrictos). Otra buena razón para participar (tanto grupos de investigación como compañías) puede ser conseguir el corpus o sus modificaciones en caso de ediciones continuadas, ya que la generación de corpus para evaluación es bastante costosa (para anotarlos o limpiarlos de errores de nulo interés para la tarea). Con esta idea se justifica el gran número de inscripciones y el menor número de participantes en tareas como HAHA o MEDDOCAN. Finalmente hay que destacar los corpus de variantes del español en varios retos, como en FACT, IroSvA o TASS.

[5] https://es.wikipedia.org/wiki/B%C3%BAsqueda_y_recuperaci%C3%B3n_de_informaci%C3%B3n
[6] https://es.wikipedia.org/wiki/Extracci%C3%B3n_de_la_informaci%C3%B3n
[7] https://en.wikipedia.org/wiki/SemEval
[8] https://en.wikipedia.org/wiki/Semantic_analysis_(linguistics)
[9] http://web.eecs.umich.edu/~mihalcea/senseval/
[10] https://en.wikipedia.org/wiki/Word-sense_disambiguation
[11] http://www.clef-initiative.eu/
[12] http://www.multimediaeval.org/

A continuación, se incluye un apartado con los resúmenes y preguntas (con sus respuestas) sobre aspectos interesantes de los nueve retos. Finalmente se detallan resultados sobre la actividad colaborativa y algunos comentarios finales.

2. Los retos de IberLEF

Se describen brevemente los aspectos más interesantes de cada uno de los nueve retos (su objetivo, organización y corpus) y se aportan algunos detalles sobre las aproximaciones que obtuvieron mejores resultados. Además de ofrecer ese resumen de la tarea, reto o *track*, las preguntas incluidas y sus respuestas muestran los aspectos que fueron especialmente relevantes para los estudiantes del máster (en general, informáticos con conocimientos amplios en tecnologías del lenguaje y no necesariamente en semántica). Además, se han incluido explicaciones básicas y referencias.

2.1 eHealth-KD

Objetivo: La tarea de descubrimiento de conocimiento en documentos de salud en español (eHealth-KD), propone la identificación de cláusulas relevantes (una cláusula es una o más palabras consecutivas) y su clasificación (en concepto, acción, predicado o referencia) y la identificación de las relaciones semánticas entre ellas (generales, contextuales, acción o predicado).

Organización: Se plantean dos subtareas distintas, la subtarea A es la identificación de frases clave y la subtarea B implica la extracción de relaciones semánticas entre dichas frases clave. En ambas se pueden emplear planteamientos basados en conocimiento y recursos de ámbito general o específicos (generalmente escasos y no anotados), modelos de *deep learning*[13], *word embeddings*[14] (pre-entrenados o propios) y otros.

Corpus: Formado por 1.045 oraciones en español de dominio médico. De ellas, 600 se emplean para entrenamiento y 100 para validación; ambos conjuntos constan de anotaciones, al igual que las 300 oraciones de la prueba[15].

Detalles sobre las aproximaciones utilizadas [43]: Participaron diez sistemas, en su mayoría empleando técnicas de *deep learning* como LSTM[16], BiLSTM[17] con *embeddings* por palabra o carácter, modelos pre-entrenados (*FastText*[18]) o entrenado en el dominio médico y CRF[19]. La solución con mejor porcentaje de acierto, del grupo de investigación TALP-UPC[20] [39], empleó BERT[21] combinado con *Grated Recurrent Units* o GRU[22] y CNN para resolver ambas tareas simultáneamente, y se entrenó con datos de ediciones anteriores del reto.

[13] https://en.wikipedia.org/wiki/Deep_learning
[14] https://en.wikipedia.org/wiki/Word_embedding
[15] El corpus o la fase de prueba también se denomina de test.
[16] https://en.wikipedia.org/wiki/Long_short-term_memory
[17] https://stackoverflow.com/questions/43035827/whats-the-difference-between-a-bidirectional-lstm-and-an-lstm
[18] https://en.wikipedia.org/wiki/FastText
[19] https://en.wikipedia.org/wiki/Conditional_random_field
[20] http://www.talp.upc.edu/
[21] https://en.wikipedia.org/wiki/BERT_(language_model)
[22] https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

La segunda solución, *coin_flipper* [6], está basada en una arquitectura conjunta de LSTMs con *embeddings preentrenados* con *FastText* y etiquetas de análisis sintáctico (POSTag[23]). Solo un participante [32] utilizó entidades Wikidata en una aproximación basada en el conocimiento combinada con una arquitectura de aprendizaje profundo.

Otra información: La principal innovación técnica del método BERT (*Bidirectional Encoder Representations from Transformers*[24]) consiste en el entrenamiento bidireccional utilizando una arquitectura *Transformer* [54] para el modelado del lenguaje, consiguiendo una mayor profundidad y riqueza en el contexto lingüístico que los modelos que son direccionales o secuenciales. La arquitectura *Transformer* contiene un mecanismo de atención[25] que promedia los pesos que pueden ser relevantes en otros puntos de una red neuronal. Este mecanismo permite incluir características lingüísticas o sintácticas al proceso de aprendizaje de la red neuronal que implementa el algoritmo del aprendizaje profundo y que aprende relaciones contextuales entre las palabras en un texto.

Los *Transformers* incluyen dos mecanismos separados: un codificador que lee el texto de entrada y un codificador que produce una predicción para la tarea. A diferencia de los modelos direccionales, que leen la entrada del texto secuencialmente, el *Transformer* lee toda la secuencia de palabras de una vez, por eso se considera bidireccional, permitiendo al modelo aprender el contexto de una palabra basándose en el resto del entorno.

Otro método muy utilizado es ULMFiT[26] (*Universal Language Model Fine-tunning for Text Classification*), que ha sido la técnica por excelencia de procesamiento de lenguaje natural hasta que apareció BERT en 2018. Consiste en un modelado del lenguaje[27], es decir en una distribución de probabilidad sobre una secuencia de palabras. Se basa en tres etapas principales: pre-entrenamiento, *fine-tunning* o limpieza del corpus y clasificación. Este método es universal porque usa una sola arquitectura y proceso de entrenamiento, requiere un pre-procesamiento no customizado y no requiere documentos adicionales en el dominio.

Preguntas planteadas en el foro:

1.1 Pregunta de E12: ¿Por qué se obtiene mejor porcentaje de acierto para todos los sistemas en ciertas extracciones como *concept* y relaciones como *target*?

Respuesta de E1: La distribución de la frecuencia de ejemplos anotados en el conjunto de entrenamiento afecta al éxito de la tarea de extracción de forma correlacionada en un 81,1 %, es decir, en gran parte de los casos cuanto mayor es el número de instancias con las que se ha entrenado (y por lo tanto aprendido), mayor es el número de instancias detectadas en el conjunto de test. Se puede observar mejor el efecto en la figura 1.

Por ejemplo, se observa que hay un mayor porcentaje de acierto para la extracción *Concept* o la relación *target* ya que el número de anotaciones disponibles en el entrenamiento para esas dos etiquetas es mucho mayor que para el resto: para *Concept* hay 2.381 anotaciones de

[23] https://en.wikipedia.org/wiki/Part-of-speech_tagging
[24] https://en.wikipedia.org/wiki/BERT_(language_model)
[25] https://cambiodigital-ol.com/2019/06/todo-sobre-el-aprendizaje-profundo/
[26] https://towardsdatascience.com/understanding-language-modelling-nlp-part-1-ulmfit-b557a63a672b
[27] https://www.kaggle.com/karthik7395/ulmfit-language-model-state-of-the-art

entrenamiento frente a las 976 de la segunda frase clave, y para *target* hay 974 frente a 511 de la siguiente relación.

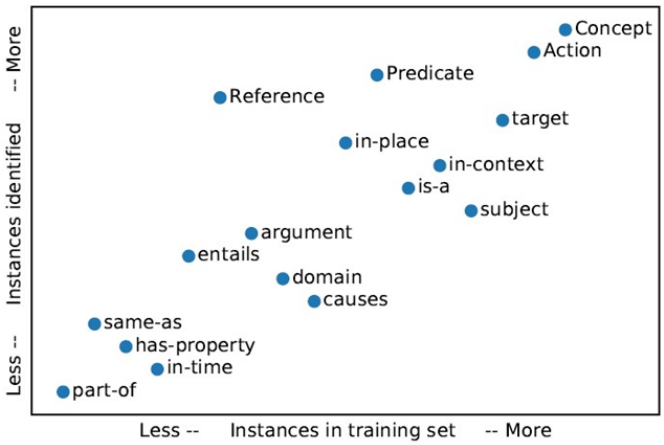

Figura 1: Número de instancias presentes en el conjunto de entrenamiento con respecto al número de instancias identificadas (extraída de [40])

1.2 Pregunta segunda de E12: ¿Por qué se opina que el uso de *word embedding* público pre-entrenado es mejor que un *word embedding* propio entrenado en el dominio específico?

Respuesta de E2: Los *word embedding* pre-entrenados y públicos, al estar entrenados en ámbitos y conjuntos de datos masivos generales (Wikipedia, Google News, etc.) permiten representar y obtener mayor cantidad de relaciones lineales entre palabras y una detección más eficiente de sinónimos. Además, los *word embedding* obtienen sinónimos por contexto de forma no supervisada, con la idea de que palabras similares ocurren en contextos similares.

Sin embargo, aunque serían deseables, los *word embeddings* entrenados en dominios específicos como en el dominio de biomedicina en español no suelen estar disponibles, principalmente porque para entrenar un modelo de embeddings hace falta, muchos datos y una alta capacidad de cómputo (grandes servidores, GPUs/TPUs, …). Por ello, también en los dominios específicos se usan modelos pre-entrenados de dominio general.

1.3 Pregunta de E1: ¿Por qué se sugiere hacer futuros experimentos para investigar el rendimiento de los sistemas *end-to-end*? Entendiéndose por *end-to-end* al sistema que aborda el entrenamiento de las dos subtareas propuestas de forma simultánea.

Un buen sistema de NER (subtarea A) retroalimenta a un buen sistema de detección de relaciones (subtarea B). Abordar ambas tareas como una sola permite escalar mejor el problema y además mejorar los resultados. Pero requiere mas investigación en el tema y para el español.

Respuesta de E5 y E12: Para comprobar que el uso de sistemas *end-to-end* mejora el rendimiento, solo se tiene los resultados de un equipo (TALP-UPC) y por lo tanto no queda claro si la ventaja procede de un mejor modelo o de un entrenamiento conjunto de las dos subtareas. Y eso a pesar de que el equipo ha obtenido mayor eficiencia que el resto de los sistemas (mayor valor F1[28]) en ambas tareas. Este equipo modela la dependencia entre las etiquetas asignadas a las frases clave y las relaciones entre ellas utilizando una aproximación basada en BERT, GRU, y datos de entrenamiento de ediciones pasadas… El resto de las

[28] https://en.wikipedia.org/wiki/F1_score

soluciones aplican modelos de *deep learning* y combinaciones de técnicas diferentes: modelos secuenciales (LSTM, biLSTM...), CNN[29], *word embeddings* (propio o pre-entrenado...), etc.

2.2 FACT

Objetivo: Análisis de la eventualidad, habitualmente llamada factualidad[30] y su clasificación (FACT), esto es, identificar hechos o eventos categorizandolos en *Fact* (F) los hechos que ocurrieron, *ContraFactual* (CF) los hechos que nunca ocurrieron y *Undefined* (U) los hechos indefinidos (refiriéndose a hipótesis, predicciones o posibilidades en el futuro).

Según Sauri (2008) [45 y 46], la tarea FACT consiste en determinar el estado de los eventos verbales con respecto a la factibilidad de los textos. La factualidad puede entenderse como la categoría que determina el estado de realización de los acontecimientos, es decir, si los acontecimientos son o no ciertos. Sin embargo, evaluar la veracidad de los hechos no forma parte de la tarea, ya que solo se solicitó el análisis de su forma (estructura). Así, esta tarea podría ser considerada como una tarea básica y previa a otras tareas más complejas como la comprobación de hechos (*fact-checking*) o la detección de noticias falsas (*fake-news*).

Corpus: El Corpus contiene textos en español (de España, es-ES, y de Uruguay, es-UY) con más de 5000 eventos verbales (el corpus inicial de textos uruguayos con 2080 eventos se amplió con textos procedentes de España). Se dividió en dos grupos, 80% para entrenamiento y 20% para prueba y contiene textos de registro periodístico, la mayor parte de ellos procedentes de las secciones políticas de periódicos españoles y uruguayos.

Participación: El número inicial de equipos participantes fue de cinco, pero un equipo fue desclasificado debido a que no entregó la descripción de su sistema, aun cuando el resultado obtenido sí fue mostrado en la tabla de resultados.

Detalles sobre las aproximaciones utilizadas [1]: los dos primeros clasificados fueron Amrita CEN [44] y Aspie96 [19]. A pesar de que el segundo clasificado también usó técnicas de *Deep learning* (*recurrent convolutional neural network*), sus resultados fueron mejorados usando *Random Forest*[31] y *Word Embeddings*, probablemente porque no utilizó una cantidad suficiente de datos para el entrenamiento del modelo, ya que los métodos de *Deep Learning* son más precisos cuando son entrenados con una mayor cantidad de datos.

Amrita CEN tenía en cuenta las diferencias en el número de apariciones de las diferentes etiquetas factuales en el corpus, asignando un mayor peso a las etiquetas minoritarias que a las mayoritarias. Aspie 96 utilizaba la tokenización para clasificar individualmente las palabras en el texto y cada palabra era representada con una serie de vectores con un indicador de si la palabra era un evento o no. Finalmente, aplicaron una matriz para clasificar cada una de las palabras en una de las tres clases indicadas.

Hay que destacar que la aplicación de *deep learning* en el reto no fue una ventaja y el grupo que mejores resultados obtuvo no lo aplicaba. Parece que la diferencia de los resultados la marcó el hecho de que el sistema de Amrita CEN compensa la menor aparición de las etiquetas minoritarias con un mayor peso para la mismas, de manera que el clasificador tiene una capacidad mayor para su reconocimiento.

[29] https://en.wikipedia.org/wiki/Convolutional_neural_network
[30] Anglicismo, palabra no recogida en el diccionario de la RAE
[31] https://en.wikipedia.org/wiki/Random_forest

Otra información: Como medida de comparación utilizaron *Macro-F1* (para combinar varias medidas de distintos clasificadores[32]), teniendo en cuenta otros factores como la *precision,* el *recall,* el *F1 score* y *accuracy*[33]. El orden de los resultados por *Macro-F1* es el mismo que el orden según el *accuracy* logrado.

Preguntas planteadas en el foro:

2.1 Pregunta de E3: En *Factuality Analysis and Classification Task* los equipos presentados proponen modelos como BERT, redes neuronales y más, ¿Por qué el equipo que propuso un sistema basado en *Word Embedding* usando *Random Forest* obtuvo mayor precisión en las métricas evaluadas?

Respuesta de E9 y E1: La diferencia de los resultados se debe a compensar la menor frecuencia de aparición de las categorías minoritarias con un mayor peso para las mismas, de manera que el clasificador tiene una capacidad mayor para el reconocimiento de estas.

Otra de las razones por las que el sistema de *Random Forest* tuvo mejor resultado es porque como el volumen de datos de entrenamiento no era demasiado grande (sobre todo en la categoría de *Counterfactual*), se perjudicaba especialmente a los sistemas de *Deep Learning*. Resultaría interesante comprobar si, una vez que crezca el número de anotaciones sobre factualidad, y con ello los corpus usados, los sistemas de *Deep Learning* igualan o sobrepasan a otras aproximaciones.

2.2 Pregunta segunda de E3: ¿Cuántas categorías se decidió utilizar finalmente para el Corpus de FACT? ¿por qué?

Respuesta de E7: Inicialmente se plantearon seis categorías, de las cuales cuatro representaban un alto grado de certeza, pero solo dos de ellas realmente denotaban si un hecho ocurría o no ocurría, objeto de evaluación; además, su reconocimiento sería extremadamente difícil.

2.3 Pregunta de E10: ¿Qué papel podría desempeñar la tarea FACT en sistemas de *fact-checking* (*fake-news*, *political ads*, etc.)?

Respuestas de E7, E9, E8, E14 y E1: Para "*fake news detection*", podría ser una parte indispensable ya que saber cómo han sido representados los eventos sería necesario en análisis posteriores. Por ejemplo, un sistema de *fake-news detection* de noticias en medios de comunicación online podría estar formado por un proceso que inicialmente anotara los eventos (*fact tagging*), otro que los categorizara (como la tarea de FACT de IberLEF2019) y otro posterior que los contrastara con una autoridad que decidiera sobre su veracidad[34].

E7 indica la existencia de un reto llamado FEVER (*Fact Extraction and VERification*)[35] en el que todavía no hay resultados suficientemente buenos [2] y reportan una precisión en la predicción de entre 65 y 70%. E14, en su primer comentario y posteriores, muestra un especial interés en la problemática que surge a la hora de usar "autoridades" fiables para contrastar la información. E10 además indica que hay una compañía llamada *Factmata* con 2 productos

[32] https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1
[33] https://en.wikipedia.org/wiki/Accuracy_and_precision
[34] https://thenewstack.io/mit-algorithm-sniffs-out-sites-dedicated-to-fake-news/
[35] http://fever.ai/

relacionados con este reto, Factmata API[36] y una extensión para el navegador, Trusted News[37]. Otra herramienta muy aconsejable para el español es la de *Meaningcloud*[38], para análisis de textos y en particular aspectos semánticos.

2.3 HAHA

Objetivo: La tarea de análisis de humor basado en anotaciones humanas (HAHA) pretende la detección y el análisis automático de texto humorístico o gracioso en lengua española (humor computacional). Se acepta que humor es la percepción de que algo es gracioso, lo que significa que la opinión de una persona lectora es determinante. Sin embargo, en un texto, se debe considerar también la intención del autor de ser gracioso o no.

Organización: El reto se divide en dos sub-tareas. La primera es de detección tweets graciosos (o humorísticos), esto es cuando la intención del autor fue ser gracioso según la opinión de jueces humanos (que lo leen). En segundo lugar, para los tuits considerados como graciosos (el 38,7% del total de tweets del corpus), los jueces humanos aportaron la "cantidad de humor" o cuánto de humorístico es el tuit, puntuándolos desde uno (intento de ser gracioso pero el autor no lo ha conseguido) hasta cinco (hilarante o muy divertido).

Para la predicción de lo gracioso que es un tweet, la subtarea fue evaluada usando *Root Mean Squared Error*[39] (raíz del error cuadrático medio) o RMSE. Se calcularon dos puntos de referencia, aleatorio y la media (pero solo se publicó el primero). Para el primer caso se escoge el valor tres para todos los tweets (mitad de la escala). La raíz del error cuadrático medio para esta referencia sobre los datos del test fue 2,455 (con la media 2,0464, la raíz del error cuadrático medio era de 1,651).

Corpus: El corpus está compuesto de 30.000 tuits anotados en español extraídos con la API *Tweepy*[40]. Fueron anotados con su cantidad de humor (11.595 humorísticos). La anotación se realizó con un esquema de votación en el que los usuarios podían seleccionar una de seis opciones: el tuit no es humorístico, o el tuit es humorístico, en cuyo caso tiene una cantidad de humor y se da una puntuación entre uno (no divertido) y cinco (excelente). El corpus se dividió en 80% entrenamiento y 20% prueba.

Las fuentes del corpus de la edición de 2018 fueron 50 cuentas de Twitter, y para la edición 2019 se añadieron todos los tweets de trece nuevas cuentas de dialecto en español, por lo que hay un total de 10.000 nuevos tweets. El conjunto de entrenamiento contiene las particiones de entrenamiento y prueba del año pasado, y algunos tweets nuevos para hacer un total de 24.000 tweets. La nueva partición de prueba consta de 6.000 tweets completamente nuevos.

El corpus de 2018 incluía algunos tweets duplicados, o tweets semánticamente muy cercanos (diferentes en unas pocas palabras) y fueron suprimidos. Esta es una mejora muy importante, porque garantiza la eficiencia del trabajo y reduce el error experimental.

Participación: 18 equipos de los 101 que se inscribieron (se supone que para obtener el corpus anotado) para la tarea 1 y trece para la tarea 2.

---

[36] https://factmata.com/api.html (acceso 17/12/2019)
[37] https://trusted-news.com/ (acceso 17/12/2019)
[38] https://www.meaningcloud.com/es
[39] https://en.wikipedia.org/wiki/Root-mean-square_deviation
[40] https://www.tweepy.org/

Detalles sobre las aproximaciones utilizadas [9]: El participante adilism [25] (el mejor en las dos sub-tareas), utilizó un modelo pre-entrenado basado en BERT multilingüe (*multilingual cased BERT-base pretrained model*)[41] junto a la librería FASTAI[42]. En primer lugar, utilizaron el conjunto de datos (*dataset*) proporcionado en la competición no anotado, esto es sin etiquetas (no supervisado[43]) para entrenar un modelo de lenguaje no supervisado basado en BERT. A continuación, lo ajustaron para cada una de las dos sub-tareas con *one-cycle learning-rate style*[44] y *discriminative fine-tuning*[45]. A partir de este momento el sistema ya pasa a ser supervisado[46] porque las técnicas de *fine-tunning* consisten en "enseñar" a los modelos pre-entrenados BERT una tarea específica, y para ello, se entrenan con las etiquetas de la tarea. Aparte, se utilizó un modelo *Naive Bayes*[47] multinomial binario con unigramas y bigramas con TF-IDF[48] (*Term frequency – Inverse document frequency*), y combinaron sus predicciones con las procedentes de la red neuronal (BERT) con una regresión logística[49] para obtener las predicciones finales. En la sub-tarea 2 cambiaron el modelo BERT para utilizar la perdida cuadrática media (*mean-squared loss*[50]) y combinar las predicciones con un modelo de árbol de gradiente (*gradient-boosted tree model*[51]) de LightGBM[52].

Kevin & Hiromi[53] construyeron un sistema basado en cinco modelos en la tarea de clasificación, combinando las predicciones con una regresión lineal y utilizando un gráfico con un límite de decisión para decidir en qué punto se maximizaba la medida F1. Los dos primeros modelos fueron pre-entrenados con 500.000 nuevos tuits. Para la sub-tarea 2 utilizaron el mismo conjunto de modelos sin el NBSVM (*Naïve Bayes - Support Vector Machine*[54]) y como salida únicamente la medida F1.

En cuanto a los resultados de la tarea HAHA, parece que apoyarse en varios modelos es lo que ha funcionado mejor. Modelos de lenguaje pre-entrenados como BERT y ULMFiT[55] fueron utilizados por los mejores sistemas. Además, el aprendizaje multitarea (MTL, *multi-task learning*) fue útil para varios equipos (beneficiandose de los resultados de una sub-tarea en la otra). En general, para aprovechar las múltiples técnicas, los participantes han construido agrupaciones (*ensembles*) como, por ejemplo, una red neuronal con un modelo Naive Bayes para mejorar sus resultados.

[41] https://github.com/google-research/bert/blob/master/multilingual.md
[42] https://docs.fast.ai/
[43] https://es.wikipedia.org/wiki/Aprendizaje_no_supervisado
[44] https://towardsdatascience.com/finding-good-learning-rate-and-the-one-cycle-policy-7159fe1db5d6
[45] https://github.com/comicencyclo/TransferLearning_DiscriminativeFineTuning/wiki/Setting-up-Discriminative-Fine-Tuning
[46] https://es.wikipedia.org/wiki/Aprendizaje_supervisado
[47] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[48] https://en.wikipedia.org/wiki/Tf%E2%80%93idf
[49] https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica
[50] https://en.wikipedia.org/wiki/Mean_squared_error
[51] https://en.wikipedia.org/wiki/Gradient_boosting
[52] https://lightgbm.readthedocs.io/en/latest/
[53] Blog de http://kevinbird15.com/2019/06/26/High-Level-Haha-Architecture.html
[54] https://medium.com/towards-artificial-intelligence/naive-bayes-support-vector-machine-svm-art-of-state-results-hands-on-guide-using-fast-ai-13b5d9bea3b2
[55] https://arxiv.org/abs/1801.06146

Preguntas planteadas en el foro:

3.1 Pregunta de E14: Analizando los resultados parece que la utilización de modelos de lenguaje pre-entrenados (*pre-trained language models*) como BERT o ULMFiT es la técnica que mejores resultados ha obtenido. No obstante ¿qué otras técnicas o mejoras han sido utilizadas para asegurar un mejor resultado?

Respuesta de E7: BERTo ULMFiT también se han combinado con el método "*Binarized Multinominal Naïve Bayes*" que es una aproximación del modelo *Naive Bayes* y se utiliza principalmente cuando las frecuencias de las palabras no juegan un papel fundamental en el estudio, como por ejemplo en el análisis de sentimientos.

También hay que destacar el uso de SVM (*Support Vector Machine*) que consiste en modelos de aprendizaje supervisados que se basan en algoritmos de aprendizaje que analizan la información para su clasificación y análisis de regresión.

Respuesta de E1: E7 ya ha comentado algunos métodos con los que se ha combinado los modelos pre-entrenados mencionados, así que ahora se comentan algunas técnicas para entrenar estos mismos modelos de forma que sean más eficientes.

El problema de los modelos pre-entrenados es que se puede caer en el olvido catastrófico[56] (es decir, que los datos usados en el entreno tengan mucho más valor que los usados en el pre-entreno "olvidando" el pre-aprendizaje) o que se produzca un sobre-entrenamiento (*overfitting*[57]). Para evitar esto se pueden usar técnicas cómo: *slanted triangular learning rate*[58] (STLR) o *gradual unfreezing*[59] que son técnicas que modifican el valor o el peso de los datos en el aprendizaje (el *learning rate* en definitiva) para mantener la relevancia de los *datasets* usados al principio. En esta tarea, los mejores equipos aplicaron esas técnicas al entrenamiento de sus modelos).

Resumen de E14: Además de técnicas de *Deep Learning*, otras mejoras han sido o continuar con el entrenamiento del modelo de lenguaje con datos procedentes de los competidores o con nuevos tweets, o bien la utilización de técnicas de aprendizaje multitarea (MTL, *multi-task learning* [2]) incluyendo el beneficiarse de los resultados de una sub-tarea en la otra.

3.2 Pregunta de E7: ¿Qué papel juega la subjetividad del autor en el análisis de humor en textos y sentimientos en general?

Respuesta de E13: Tanto los autores de los mensajes, como los que después los analizan, aportan su propia opinión, con lo que la subjetividad tiene un papel muy importante. Y también en general cuando se realiza un análisis de otros sentimientos, como la felicidad o el odio y se necesita entender o, mejor aún, conocer la intención del autor.

Además, los resultados se ven también condicionados por la subjetividad de las personas que clasificaron tweets, ya que no hay un estándar establecido para determinar si algo es humorístico o no, y en qué grado. Pensar que las 800 personas que participaron tienen la misma idea de ese tema es algo arriesgado, y el método de clasificación a pesar de requerir de

[56] https://en.wikipedia.org/wiki/Catastrophic_interference
[57] https://en.wikipedia.org/wiki/Overfitting
[58] https://towardsdatascience.com/multi-task-learning-in-language-model-for-text-classification-c3acc1fedd89
[59] https://github.com/huggingface/transformers/issues/674

varias "opiniones", sigue siendo subjetivo. Para esta tarea o retos, sería importante tener en cuenta la ironía y establecer mecanismos de tratamiento y reconocimiento de ésta última.

2.4 IroSvA

Objetivo: Tarea (IroSvA) que consiste en detectar la ironía en mensajes escritos en las variantes de español de España, México y Cuba.

Organización: Se dividió en tres subtareas correspondientes a las variantes de español mencionadas. Participaron doce equipos.

Detalles sobre las aproximaciones utilizadas [41]: Principalmente se utilizaron métodos de aprendizaje automático: desde máquinas de soporte vectorial (*SVM*) a redes neuronales con modelos como *Word2Vec*, *FastText*, *Doc2Vec*, *Elmo*, *Bert* y *n-gramas* a nivel de caracteres y palabras.

El equipo mejor clasificado, EliRF-UPV [23], logró una F1-promedio de 0,6832 con una solución basada en aprendizaje profundo, que calcula representaciones vectoriales combinando la etapa de codificación de un modelo transformador con *word embeddings* extraídos a partir de un modelo de *skip-grammar* entrenado mediante *Word2Vec* con 87 millones de *tweets*.

El equipo CIMAT [40] fue el segundo clasificado con una F1-promedio de 0,6585. Su sistema forma vectores mediante la concatenación de características obtenidas a partir de tres representaciones distintas: (i) *word embeddings* generados por *Word2Vec* en grandes corpus, (ii) con base en una representación profunda (*deep representation*) apoyada por redes neuronales *LSTM*, y (iii) *n- gramas* a nivel de carácter y de palabra. Todas las representaciones alimentan un *SVM* con un núcleo lineal.

Sólo tres aproximaciones tuvieron en cuenta el contexto del mensaje para detectar la ironía. De éstas, la que obtuvo mejores resultados quedó quinta en la clasificación global. Ninguna de las tres superó el indicador *baseline-w2v* establecido por la organización. Con lo que, en general, no existe evidencia de que tener en cuenta el contexto mejore la detección de la ironía en mensajes cortos.

Preguntas planteadas en el foro:

4.1 Pregunta de E5: ¿Cómo afecta en la detección de la ironía tener en cuenta el contexto del mensaje?

Respuesta de E7: En una tarea de detección de la ironía en textos, al igual que en el análisis de sentimientos, parece que debería ser fundamental tener en cuenta el contexto. Como la ironía es algo subjetivo y está ligado a la opinión del autor, es necesario utilizar mecanismos y métodos que permitan explorar el entorno en su conjunto e intentar predecir si el autor tiene o no una intención irónica.

Además, hay muchos tipos de ironía y para otorgar el correcto valor semántico de cada elemento comunicativo habrá que saber distinguirlos. Hay ironías que se centran en la contradicción de sentimientos. En este caso, interesa fijarse si se usan palabras o expresiones con contenido semántico positivo o negativo para decir justo lo contrario. Por ejemplo, en una página en la que todo participante cuenta historias negativas que le han ocurrido como en

"Asco de Vida", hay que dotar al sistema de una etiqueta previa de polaridad[60] negativa que tenga peso al computar la polaridad global. Se puede deducir que hay ironía si un usuario escribe "Me alegró el día" al final de una historia, conociendo su intención de transmitir un mensaje negativo. Se podría entrenar el sistema para detectar ironías de este tipo.

En otros casos, hay una contradicción entre lo que se dice (proposición) y a lo que se refiere (referente) como, por ejemplo, "¡Qué grande eres niño!" y el niño no llega ni al metro de altura. En este contexto habría que entrar a valorar antónimos.

Finalmente, hay que darse cuenta de que cuando los participantes intentan detectar la ironía no tratan de entender los mensajes, sino construir un clasificador binario que determine si la representación de la semántica de un mensaje irónico lo indica.

2.5 NEGES

Objetivo: La tarea relacionada con la negación en castellano (NEGES), trata este fenómeno lingüístico como una fase necesaria en posteriores análisis de sentimiento o recuperación de información.

Organización: Se plantea como la resolución de dos subtareas. La subtarea A para la detección de cláusulas y sentencias negativas (consecutivas o no): *no*, *sin*, *ni siquiera*, *sin ningún*, *no ... apenas*, *no ... nada*, etc. La segunda subtarea (B), se propone para evaluar el impacto de la negación en el análisis de sentimiento, clasificando cada opinión como positiva o negativa.

Para las métricas de evaluación se contabilizan los casos correctos (*True Positive*), es decir, la identificación correcta de las partículas de negación, y para casos correctos parciales se contabilizan los *False Negative* y se calculan las métricas de precisión, cobertura y medida F1.

Corpus: Conjunto de documentos de opiniones de películas, libros y productos en español. Se utiliza el corpus SFU Review_{SP}-NEG [27], para entrenar y testear los sistemas tanto para la Subtarea A (formato CoNLL[61]), como para la Subtarea B (formato XML[62]).

Participación: Participaron trece equipos, de los cuales enviaron soluciones cinco de ellos, cuatro para la subtarea A y un solo participante para la tarea B.

Detalles sobre las aproximaciones utilizadas [26]: En la subtarea A, el equipo con mejor evaluación, CLiC [4], emplea un modelo basado en *Conditional Random Field,* o MRF (*Markov Random Field*). Este equipo utilizó como características las formas de las palabras y las etiquetas PoS de la palabra actual, la palabra posterior y las seis palabras previas (contexto). El equipo de la UNED [14] presenta una solución basada en un modelo neuronal profundo con BiLSTM*, embedding* por carácter y etiquetas sintácticas (*POSTagging*) en forma de vector unitario por etiqueta (*one-hot*), con un post-procesado con reglas que mejoran la eficiencia, al corregir errores frecuentes cometidos por la red.

Para la Subtarea B, hubo un solo participante: LTG-Oslo [3]. Este equipo utilizó un enfoque de aprendizaje multitarea en el que un modelo único se entrena simultáneamente para ambas subtareas, una red neuronal BiLSTM, donde las capas bajas se usan para predecir la negación y las capas altas para predecir el sentimiento, a nivel de oración.

[60] https://es.wikipedia.org/wiki/An%C3%A1lisis_de_sentimiento
[61] http://mwetoolkit.sourceforge.net/PHITE.php?sitesig=MWE&page=MWE_070_File_types&subpage=MWE_010_CONLL
[62] https://en.wikipedia.org/wiki/XML

Preguntas planteadas en el foro:

5.1 Pregunta de E8 y E12: La edición de 2019 de la tarea NEGES consistió en dos subtareas ¿A qué puede deberse el bajo número de envíos de la segunda subtarea?

Respuesta de E12: El número bajo de envíos de la subtarea B puede deberse a que el esfuerzo y las soluciones realizadas para la subtarea A, detección de la negación, no puede ampliarse y aplicarse directamente en el análisis de sentimiento.

Respuesta de E3: Para entender el efecto de la primera tarea en la segunda ("El papel de la negación en el análisis de sentimientos"), es necesario representar la negación de una manera eficiente. Y por ello, seguramente, los equipos evaluando trabajo, esfuerzo y complejidad se centraron en poder realizar lo mejor posible la primera tarea, ya que debía ser eficiente antes de proseguir con la segunda tarea.

Respuesta de E4: La segunda tarea requiere de un gran trabajo previo para el tiempo que disponían los equipos. Los equipos participantes sabían que no iban a poder ofrecer unos resultados llamativos y prefirieron centrarse en la otra tarea más abarcable. Sin embargo, es loable la original solución que ofreció el equipo de Oslo.

Respuesta resumen de E8: El bajo número de envíos en la tarea 2 puede deberse a que, para estudiar el impacto de la detección precisa de la negación, en el análisis de sentimiento, es necesario determinar cómo representar eficientemente la negación (en el caso de sistemas de aprendizaje automático), o cómo modificar la polaridad de las palabras dentro del ámbito de la negación (en el caso de sistemas basados en léxico). Para la subtarea 2, incluso la aproximación del equipo LTG-Oslo, que se lleva a cabo a nivel de documento, es complicada porque en este caso es difícil determinar exactamente señales de negación (“Ya estaba casi, ¿no?”, “...a no ser que...”).

2.6 NER_Portuguese

Objetivo: Reconocimiento de Entidades Nombradas (NER) y Extracción de Relaciones (RE) para el idioma portugués.

Organización: Hay tres tareas (o subtareas).

> **Sub-tarea 1 (NER):** Reconocimiento de entidades nombradas (*Named Entity Recognition*, NE) de tipo persona, valor, fecha, hora, lugar y organización, en tres corpus muy diferentes. Uno de temas generales (SIGARRA). Otro con textos en portugués europeo y brasileño (HARLEM) de textos médicos (con entidades de tipo persona), corpus de mucho interés ya que utiliza abreviaciones médicas, jerga y caracteres especiales para los nombres. El tercer corpus o *dataset* es de fichas policiales en portugués brasileño, por lo tanto, con documentos bien estructurados.
>
> **Sub-tarea 2 (RE):** Extracción de relaciones entre NE, solo de tipo persona, lugar y organización. En este caso se podía usar un conjunto de test que ya tuviera anotadas las NE u otro sin estas anotaciones. El *dataset* de portugués utilizado era combinación de dos corpus *Golden Collections from HAREM conferences* y *Summ-it++* [16].
>
> **Sub-tarea 3 (OIE):** *The open relation extraction* es una tarea más general que la anterior. Hay que encontrar relaciones explícitas en nombres o grupos nominales (*Noun Phrases, NP*) o grupos en los que la palabra central es un nombre y el grupo en su totalidad

representa a su vez a un nombre (sin restricciones impuestas por el tipo del NE). El *dataset* fue el *Portuguese Open IE corpus*. Había dos conjuntos de prueba (o test), uno con las NP anotadas y el otro no.

Corpus: Se ofrecían cinco conjuntos de datos en portugués (*datasets)* para la evaluación de los sistemas presentados, tres para NER y dos para RE.

Participación: En la Tarea 1 (NER) participaron cinco grupos, uno de los cuales presentó dos sistemas. En la Tarea 2 (RE) solo un sistema participó en el Test 1 y ninguno en el Test 2. En la Tarea 3 dos grupos participaron en el Test 1, pero presentaron un total de cinco sistemas; en el Test 2 participaron los dos grupos que se presentaron en el Test 1 y un grupo más.

Detalles sobre las aproximaciones utilizadas [10]:

En cualquier lengua, la identificación de las NEs (*Named Entities*) y sus categorías (organización, lugar, persona, etc) y la extracción de relaciones entre entidades (por ejemplo, la afiliación entre una persona y una organización), son dos tareas básicas y el primer paso en el análisis semántico de un texto. Y para llevarlas a cabo es necesario disponer de recursos.

En la tarea 1, los tres mejores sistemas para estos dos conjuntos de datos usaron aproximaciones basadas en Redes Neuronales y Modelos de Lenguaje. Los resultados para el conjunto de datos policiales mostraron una notable diferencia entre las aproximaciones basadas en Redes Neuronales y las que no.

Para la Tarea 2, el sistema que participó en el Test1, y aunque no se menciona en el artículo, usa reglas basadas en etiquetas PoS (*Part-of-Speech*).

Para la Tarea 3 y el Test 1, DPTOIE [48, 49 y 50] obtuvo los mejores resultados. En el Test 2, se establecieron cuatro escenarios diferentes. En general, los sistemas DPTOIE y Linguakit [17 y 18] fueron los que rindieron mejor en todos los escenarios de evaluación. No hay detalles en el artículo sobre cómo están implementados estos sistemas, pero DPTOIE usa el analizador de dependencias de Stanford y reglas específicas para extraer hechos de frases en portugués. Linguakit, por su parte, utiliza un analizador de dependencias, un etiquetador *PoS* y otros métodos basados en *Open Information Extraction* [20].

Preguntas planteadas en el foro:

6.1 Pregunta de E11: En NER_Portuguese, en el Test 2 de la Tarea 3 se proporciona a los participantes datos sin anotaciones referentes a los grupos o sintagmas nominales. Así que los sistemas han de extraer y clasificar primero estos sintagmas de las frases y después obtener los descriptores de relación entre pares de sintagmas ('director de', 'está en', 'profesor de', ...). Como la tarea es difícil y para el portugués hay pocos recursos, al pasar a la fase de evaluación de la Tarea 3, los organizadores deciden realizarla en varios escenarios distintos (divide y vencerás). ¿Cuál es la razón por la que se contemplan estos escenarios distintos?

Respuesta de E14: La razón es que en el primer escenario (que es el más simple) en el que solo se utiliza el conjunto de datos del Test 2 para la evaluación, los sistemas son muy torpes para identificar las relaciones correctas que no están en este conjunto de datos (las métricas son bajísimas). Los valores de las métricas de evaluación del Test 2 a partir del segundo escenario aumentan considerablemente con respecto al primero. La razón de plantear nuevos escenarios

podría ser salvar la visibilidad positiva de las métricas y levantar el ánimo de los participantes siendo menos restrictivos y evidenciando la potencialidad del trabajo hecho.

Respuesta de E11: Sí, el primer escenario presenta ese problema: un sistema podría estar preparado para extraer relaciones correctas, que al no estar en el corpus de prueba, no permitiría valorar bien la precisión del sistema.

Pero los escenarios 2, 3 y 4 tienen otras particularidades que podrían afectar a la precisión, a la exhaustividad (*recall*) o a ambas. Resulta que los conjuntos de datos de entrenamiento y prueba están compuestos de decenas de relaciones elegidas aleatoriamente de entre 25 frases en portugués. Parece que el origen del problema estaría en que el conjunto de datos de entrenamiento y los dos de prueba son disjuntos entre sí, así que a la hora de evaluar la Tarea 3 se ha de considerar la posibilidad de que un sistema identifique correctamente una relación que no esté en el conjunto de datos de prueba considerado. Si no se hace, se podrían obtener métricas sesgadas tanto para la precisión como para la exhaustividad. Pero, por otra parte, ¿qué sentido tiene evaluar en el mismo corpus de entrenamiento? Parece que deberían tenerse buenísimos resultados, si se ha aprendido bien.

Luego, la posibilidad de obtener métricas sesgadas es lo que hace que los organizadores propongan pruebas o evaluaciones en distintos escenarios. En cada escenario se consideran distintas relaciones como objetivo (realizando distintas combinaciones entre las relaciones provenientes del conjunto de datos de entrenamiento y de los dos de prueba). Después comparan las métricas obtenidas en cada escenario, y como estas son consistentes en los cuatro escenarios, llegan a la conclusión de que los resultados obtenidos son “robustos y fiables.”

6.2 Pregunta de E2: Referida a la primera tarea en la que se tiene que extraer NE. En la tarea se proporcionan tres conjuntos de textos para la prueba los sistemas presentados; el general, el policial (textos oficiales de la policía de brasileña), y el clínico (textos creados por médicos, enfermeros, etc.). ¿Cuál podría ser un motivo por el que los resultados obtenidos para el conjunto policial son muy superiores a los de los otros dos?

Respuesta de E4: No se puede acceder al corpus clínico ni al policial por la vulnerabilidad del contenido, pero en [10] indican las siguientes características de los corpora. El corpus de la policía de Brasil se caracteriza por estar muy bien estructurado, lo que ayuda al pre-procesado y a mantener la estructura del texto tras haberlo procesado. Tampoco tiene errores gramaticales, como sí el corpus clínico, lo que ayuda a detectar correctamente las entidades de tipo *Person*. El corpus clínico tiene además muchas menos frases de entrenamiento (el corpus médico cuenta con 77 *NE*, mientras que el policial cuenta con 916 para generalizar). Por lo tanto, el corpus policial por su especificidad y riqueza favorecen el entrenamiento de una red neuronal y se obtienen buenos resultados.

Resumen de E2: Como indica E4, el corpus policial está formado por textos bien estructurados en los que predomina la calidad gramatical. En el otro extremo, el corpus clínico está caracterizado por el uso de abreviaciones y de tecnicismos.

Estas diferencias, unidas a que la única técnica de pre-procesado fue la tokenización de los textos, hacen que sea mucho más difícil encontrar una estructura y un patrón a los modelos presentados para extraer (NE) del conjunto de textos médicos, por mucho que hayan sido pre-entrenados los modelos. Es posible que el uso de otras técnicas de pre-procesado como

stemming o *lemmatization*[63] pudieran mejorar los resultados del conjunto de documentos médicos al reducir las palabras a una raíz, evitando errores gramaticales o depender menos de la estructura. Pero como en la tarea se explicitaba que tanto AnaR1 como *####Paulo* se consideraban de tipo persona, el uso de las anteriores técnicas no era procedente debido a que esos nombres forman parte del lenguaje en un contexto médico. De hecho, por este motivo, en un momento dado los organizadores alertaron a los participantes de que algunos modelos estaban extrayendo tokens acortados, o incluso ignorándolos.

2.7 MEX-A3T

Objetivo: La tarea de detección de autoría y agresividad en Twitter en español de México (MEX-A3T) pretende fomentar el análisis del contenido de la red social en la variante mexicana del español, ya que esta variante es bastante diferente al español europeo.

Organización: Para conseguir el objetivo presenta dos tareas: (1) Perfilado de autores (*author profiling*) y (2) Detección de agresividad en los tweets.

Corpus: El idioma del corpus es el español mexicano.

Participación: Se presentaron ocho equipos, de los cuales dos participaron en la tarea de perfilado de autores y seis en la de detección de agresividad en los tweets.

Detalles sobre las aproximaciones utilizadas [13]:

Para la tarea de perfilado de autores el equipo con la mejor aproximación fue Cerpamid [8], con vocabularios específicos para cada perfil, construidos basándose en el método *Transition Point*. A continuación, interpretan los tweets como documentos de los que extraen una serie de tokens y que posteriormente utilizan para representar al tweet. Por último, la fase de clasificación se realiza con una máquina de vector soporte (SVM) implementada en Weka[64].

Para la tarea de detección de agresividad en los tweets el mejor resultado fue el del participante *UACh* [7]. En su artículo de descripción del sistema sí que mencionan que hicieron una fase de pre-procesado de la información. Hacen uso de n-gramas de palabras y caracteres (facilitan una tabla con los n-gramas más representativos), y de *Document Embeddings*[65]. Para la clasificación utilizaron una máquina de vector soporte (SVM) y un perceptrón multicapa[66]. Este último método fue el elegido como sistema final de clasificación. Pretendía dar un contexto a los tweets y después afrontar la tarea, pero esa contextualización no produjo mejoras en los resultados finales ya que los mejores resultados los obtuvieron cuando utilizaron sólo las características léxicas mencionadas.

Preguntas planteadas en el foro:

7.1 Pregunta de E13: En la segunda tarea, hubo varios tweets que no fueron correctamente clasificados por ningún sistema. Todos estos tweets eran agresivos y fueron clasificados como no agresivos. ¿A qué se debe este hecho y cómo podría mitigarse o solucionarse en análisis de este tipo?

[63] https://www.datacamp.com/community/tutorials/stemming-lemmatization-python (acceso 17/12/2019)
[64] https://sourceforge.net/projects/weka/
[65] https://towardsdatascience.com/document-embedding-techniques-fed3e7a6a25d
[66] https://en.wikipedia.org/wiki/Multilayer_perceptron

Respuesta de E5: Los tweets no detectados usaban ironía, palabras no pertenecientes al vocabulario de entrenamiento, entidades nombradas y ofensas sin utilizar palabras vulgares. Habría que estudiar si ampliar el vocabulario de entrenamiento y utilizar métodos empleados en detección de ironía, mejoraría o no la clasificación de tweets agresivos.

2.8 TASS

Objetivo: Análisis de sentimiento en tweets en español (TASS)[67].

Organización: Incluye dos subtareas:

1. Análisis de sentimiento monolingüe: El objetivo de esta tarea es la evaluación de la polaridad de los sistemas de clasificación, a nivel de tweet, para tweets escritos en español.
2. Análisis de sentimiento "*crosslingual*" con diferentes subconjuntos del corpus InterTASS. El objetivo de esta tarea es similar al de la tarea anterior, pero los sistemas deben ser entrenados con una o más variantes del español y se prueba con una variante española diferente. La variante española del conjunto de entrenamiento tiene que ser diferente de la de evaluación, con el fin de probar la dependencia de los sistemas de un idioma.

Corpus: El *corpus Internacional TASS* fue lanzado por primera vez en la edición de 2017, y solo se componía de tweets escritos en el lenguaje español utilizado en España. La segunda versión de InterTASS es de la edición de 2018, y se agregaron un conjunto de tweets escritos en lenguaje español utilizado en Perú y Costa Rica. En la edición de 2019 se incorporaron los tweets escritos en las variantes del lenguaje español de México y Uruguay.

Por lo tanto, esta última versión contiene los tweets escritos en cinco variantes diferentes de español de: España, Perú, Costa Rica, Uruguay y México. Cada tweet fue etiquetado como positivo (P), neutral (NEU), negativo (N) y sin etiqueta de sentimiento (NONE).

- Datos en español: 3.401 tweets (P: 1.104, NEU: 418, N: 1.404, NONE: 475).
- Variante costarricense: 2.363 tweets (P: 707, NEU: 297, N: 912, NONE: 447).
- Variante peruana: 3.005 tweets (P: 756, NEU: 701, N: 820, NONE: 728).
- Variante uruguaya: 2.857 tweets (P: 912, NEU: 572, N: 1.146, NONE: 227).
- Variante mexicana: 3.000 tweets (P: 997, NEU: 249, N: 1.502, NONE: 252).

Participación: Siete equipos de investigación presentaron resultados para la subtarea 1, y cuatro equipos para la subtarea 2.

Detalles sobre las aproximaciones utilizadas [12]: Para la primera tarea, teniendo en cuenta los resultados obtenidos para cada una de las variantes de español, el equipo ELiRF-UPV [22] fue el que obtuvo, en general, los mejores resultados. El equipo propuso un sistema enfocado principalmente a emplear los codificadores del modelo *Transformer*, basado en mecanismos de atención [33], y prescinde de convoluciones y recurrencias para aprender relaciones de rango largo. Solo usan la parte del codificador para extraer representaciones vectoriales que son útiles para realizar el análisis de sentimiento. Denotan esta parte de codificación del modelo *Transformer* como "*Transformer Encoder*".

[67] https://www.w3.org/community/sentiment/

Para la segunda tarea, en tres de los cinco experimentos, el equipo Atalaya [35] obtuvo los mejores resultados, siendo el segundo en la evaluación del español de Costa Rica. Los valores obtenidos en la evaluación de esta tarea son similares a los de la Tarea 1, aunque un poco más bajos, lo cual es razonable ya que no se permitieron datos de entrenamiento de la variante española objetivo.

El equipo Atalaya utiliza diferentes representaciones de datos (*bag-of-words, bag-of characters, tweets embedding*). La novedad es el uso de dos técnicas de incremento de datos para tratar su escasez: el aumento de traducción bidimensional, y una técnica que genera nuevas instancias combinando mitades de tweets.

GTH-UPM [21] presentó un sistema cuyo *score* final era la probabilidad promedia de los resultados de los tres clasificadores distintos: uno para la clasificación de polaridad en función de los vectores característica extraídos de los *tweets*, otro de base neuronal aplicado con *FastText*, y un tercero basado en redes neuronales profundas que se servía de BERT.

Otra información: Debido a la homogeneidad de los resultados entre los distintos sistemas participantes, parece que el mayor escollo en esta área de investigación es la falta de un corpus más grande, puesto que el sistema que mejor puntuación obtuvo era el que mayor esfuerzo dedicó a compensarlo, y que los resultados en el análisis "mono" no fueran mejores que los del análisis "cross".

Actualmente otro foco de atención es el impacto del contexto social en el análisis de sentimientos [52]. Las medidas de evaluación utilizadas fueron: *Macro F1*, *Macro Precision* y *Macro Recall*[68].

Preguntas planteadas en el foro:

8.1 Pregunta de E9: ¿Cómo de significativa para los resultados es la presencia o ausencia de la variedad del español del conjunto de prueba en el conjunto de entrenamiento?

Respuesta resumen de E9, con las respuestas de E11, E7, E3 y E8:

Como señala E11, con quien coincide E7, "los resultados obtenidos en la Tarea 1 (en la que los sistemas eran evaluados y entrenados con la misma variante del español) son ligeramente mejores que los que se obtuvieron para la Tarea 2 (en la que no se permite usar la misma variante del español para el entrenamiento y la evaluación)". Diferencia achacable, como señala E8 a que "cada variante de español tiene sus peculiaridades", opinión apoyada por E3.

Si se comparan los resultados de las tareas 1 y 2 para la modalidad del español de España, el mejor resultado se obtiene en la tarea "mono"; para la modalidad de Perú la mejor puntuación se da en la tarea *cross-lingual*; en la modalidad de Costa Rica, tuvo mejores resultados la tarea "mono"; con respecto a la modalidad de español de Uruguay, que la tarea "mono" obtenga mejores resultados no es tan obvio y, para el de México, se impone de nuevo la tarea mono y sus resultados con respecto a la tarea *cross-lingual*.

Por otro lado, si se analizan los resultados de aproximaciones que utilizan el mismo modelo para ambas tareas, las diferencias en *Macro F1*, *Macro Precision* y *Macro Recall* (siendo el signo negativo una diferencia a favor de la Tarea 2) se observa que, aunque la presencia de la variante del español evaluada en el corpus de entrenamiento parece mejorar los resultados de

[68] https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

los diferentes sistemas, dicha mejora no es muy significativa ni en la comparación de los mejores la Tarea 2 con el resto de los de la Tarea 2, ni en la comparación de los resultados de cada sistema en cada una de las dos tareas.

2.9 MEDDOCAN

Objetivo: Anonimización de documentos médicos en español y evaluación del rendimiento de sistemas que pretenden identificar y clasificar información confidencial en documentos de casos clínicos escritos en español y no estructurados.

Organización: Esta tarea tenía dos escenarios (o subtareas) diferentes: (1) *NER* y clasificación, esto es la identificación y clasificación de información sensible, por ejemplo, nombres de pacientes, teléfonos, direcciones, etc; (2) la segunda subtarea se centró en identificar y enmascarar datos confidenciales sensibles para la publicación de documentos clínicos anonimizados (a su vez se dividía en 2A y 2B).

Corpus: Para esta tarea se anotó manualmente un corpus de casos clínicos enriquecidos con expresiones *PHI* (*Protected Health Information*), llamado corpus MEDDOCAN formado por 1000 casos clínicos (es un corpus sintético). El corpus se dividió aleatoriamente en tres subconjuntos: el conjunto de entrenamiento con 500 casos clínicos, y los conjuntos de desarrollo y pruebas de 250 casos clínicos cada uno.

Participación: De los 51 equipos registrados, presentaron resultados 18 de ellos.

Detalles sobre las aproximaciones utilizadas [36]: Uno de los objetivos principales de la primera subtarea era desarrollar sistemas capaces de anonimizar completamente la información sensible de los documentos clínicos. Sin embargo, ninguno de los sistemas presentados logró anonimizar toda la información sensible, pero sí estuvieron cerca.

Para la subtarea 1, *NER* y clasificación del tipo de entidad, el sistema mejor clasificado fue presentado por lukas.lange [34], con un valor-F de 0,96961, estando relativamente cerca del siguiente clasificado, Fadi [24], con un valor-F de 0,96327. Prestando atención a la cobertura obtenida por los sistemas, el mejor rendimiento fue el de lukas.lange, con una cobertura de 0.96944 seguido del participante FSL [47], con una cobertura de 0,96043.

Para la subtarea 2A, el sistema de lukas.lange obtuvo la puntuación más alta, con un valor-F de 0,97491. El segundo equipo fue Fadi, con un valor-F de 0,96861. Los mejores resultados en términos de cobertura los obtuvo lukas.lange, con una cobertura de 0,97474, mhjabreel [25], con una cobertura de 0,96591, y FSL, con una cobertura de 0,96520.

Los resultados para la subtarea 2B sorprendieron a los organizadores. El sistema con puntuación más alta fue presentado por lukas.lange, con un valor-F de 0,98530, pero el segundo equipo para esta subtarea fue jiangdehuan [28], con un valor-F de 0,98184, muy cerca del anterior. Teniendo en cuenta que jiangdehuan ocupó el séptimo lugar para las subtareas 1 y 2A, este aumento en el rendimiento necesitaría un análisis más profundo. Finalmente, los mejores resultados en términos de cobertura fueron obtenidos por jiangdehuan, con una cobertura de 0,98335, lukas.lange, con una cobertura de 0,98264, y, mhjabreel, con una cobertura de 0,97471.

El sistema presentado por lukas.lange entrena redes neuronales recurrentes con capas de salida de campo aleatorio condicional (*Conditional Random Field*). Experimentaron con inserciones *FLAIR*[4] para español que posteriormente utilizaron como entrada de las redes. En

las diferentes ejecuciones, exploraron a fondo las ventajas de inserciones *FastText* específicas de dominio pre-entrenadas en artículos de SciELO[69] y Wikipedia.

Fadi Hassam emplea un conjunto de reglas creadas automáticamente (modelo *RegEx*) y campos aleatorios condicionales (modelo *CRF*) para identificar campos *PHI* en documentos médicos. También utiliza el esquema de etiquetado BIO para establecer las etiquetas de los tokens (las tres etiquetas posibles B, I, O, indican si la palabra está al principio, en medio o fuera de una entidad PHI).

Uno de los participantes de la UNED [29, 31], que ya tiene experiencia previa tanto en la extracción de información de documentos en español como en aproximaciones mixtas [5, 37, 38, 42], presentó (siguiendo la aproximación de [53 y 11]) un modelo híbrido Bi-LSTM con CNN de cuatro niveles que puede reconocer entidades multipalabra utilizando el formato BIO y mejorando el vocabulario con enfermedades, problemas de salud, tratamientos y otras entidades encontradas en *wikidata*. El sistema no requiere refinamiento manual y los resultados muestran que para el español los *wikipedia2vec pretrained embedding vectors* son mejores que otras aproximaciones, como *Fastext* o *Glove*.

Preguntas planteadas en el foro:

9.1 Pregunta de E4: En la segunda subtarea, se organizaron dos experimentos para evaluar el rendimiento de los sistemas *ensemble*. El primer experimento se basó en un sistema conjunto combinado (*joint system*), y el segundo experimento, en un sistema de votación (*voting system*) ¿por qué un *joint system* y un *voting system* sirven para evaluar el rendimiento de estos sistemas combinados para anonimizar?

Respuesta de E4: Estos dos sistemas de agrupación de recursos (*ensemble*)[70] son conocidos dentro del mundo de la investigación y han sido previamente utilizados.

3. Actividad colaborativa

Como ya se ha indicado en la introducción, este resumen es la compilación de los trabajos realizados en el marco de la asignatura "Semántica y pragmática en la web" del máster de Tecnologías de la Lengua de la UNED, en el curso 2019-20. Cada estudiante debía seleccionar un reto para plantear al menos una pregunta sobre algún aspecto de su interés y de la que supiese la respuesta. Tras leer las contestaciones recibidas en el tiempo estipulado, tenía que incluir un resumen, tanto en el foro para ser leído por el resto de los compañeros, como en un trabajo individual que incluyese sus propias conclusiones. También debían contestar en el foro del aula virtual al menos a una pregunta de otro compañero. Como intentaban no repetir el reto, no se puede concluir sobre la preferencia de los estudiantes sobre ellos (ver tabla 3), aunque HAHA y FACT parecen los preferidos.

La actividad se describía en el plan de trabajo del aula virtual como sigue:

> Acceda a http://ceur-ws.org/Vol-2421/ "Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)" para identificar los nueve "Track" o temas de interés en que se han clasificado los trabajos (eHealth-KD, FACT, HAHA, IroSvA, NEGES, NER_Portuguese, MEX-A3T, TASS, MEDDOCAN). *Seleccionar dos de las nueve tareas, y con la información recabada:*

[69] http://scielo.isciii.es/scielo.php

[70] https://www.researchgate.net/publication/230867318_Ensemble_methods_A_review

1. Elaborar un informe breve identificando en las tareas seleccionadas (leer únicamente los artículos de Overview), descargados y disponibles en la carpeta de documentos del aula virtual, subcarpeta Tema 1: (1) el objetivo de la tarea, (2) el idioma de la tarea (de los corpus utilizados), (3) el número de participantes y (4) algún comentario sobre el tipo de aproximación que utilizaban los dos primeros sistemas o grupos "ganadores" según la evaluación de los resultados indicada en el Overview.
2. Plantear una pregunta, al menos, indicando el track al que se refiere (en el hilo denominado PEC1 del foro de este tema), para que los compañeros la respondan (debe conocer la solución para poder interactuar cuando lo considere necesario en el foro).
3. Cada ponente de una pregunta resumirá y publicará en el foro la respuesta elaborada colaborativamente (incluyendo su propia respuesta). Como las preguntas no deben repetirse, se utilizará el orden cronológico en caso de conflicto, y cada uno planteará una pregunta diferente.
4. Finalmente, hay una entrega individual del documento con las respuestas anteriores y comentarios sobre la resolución colaborativa, en la sección "entrega de trabajos" Tarea PEC1.

La actividad colaborativa se realizó durante cuarenta días y en el foro del aula virtual los estudiantes participaron en conversaciones pregunta y respuestas según la tabla 1.

Track/Reto	**Acrónimo**	**N. Convers.**
Track1	eHealth-KD	3
Track2	FACT	3
Track3	HAHA	4
Track4	IroSvA	1
Track5	NEGES	1
Track6	NER_Port.	2
Track7	MEX-A3T	1
Track8	TASS	1
Track9	MEDDOCAN	1

Tabla 1: Conversaciones en los nueve *retos* de IberLEF 2019

La participación de los estudiantes en el foro (identificados por EN) fue (ver tabla 2):

	N. Intervenciones	**N. Resp.**
E1	6	4
E2	3	1
E3	4	2
E4	7	6
E5	4	2
E6	1	0
E7	10	5
E8	5	3
E9	4	2
E10	5	3
E11	3	1
E12	6	2
E13	2	1
E14	5	4

Tabla 2: Número de intervenciones de cada estudiante y de ellas, número de respuestas a preguntas de sus compañeros

En la tabla 3 se indica para cada reto de IberLEF2019, el número de veces que fue seleccionado para ser resumido y cuántas veces el estudiante hizo una pregunta (columna con) o no (columna sin). En la cuarta columna se indica el número total de preguntas. En general, las preguntas se refieren a aspectos planteados en los artículos de resumen.

	veces resumido	con preg	sin preg	Nro. Preg
eHealth-KD	2	2	0	3
FACT	7	2	5	3
HAHA	6	3	3	4
IroSvA	1	1	0	1
NEGES	3	1	2	1
NER_Port.	2	2	0	2
MEX-A3T	1	1	0	1
TASS	4	1	3	1
MEDDOCAN	2	1	1	1

Tabla 3: Datos sobre retos y preguntas de interés

4. Comentarios finales

Esta actividad se planteó aprovechando la publicación *on-line* y en abierto de las aproximaciones utilizadas en los retos del foro de evaluación IberLEF 2019, justo antes del comienzo del curso académico 2019-2020. En esta compilación se han incluido muchos detalles encontrados en estos documentos, sobre el estado actual de la investigación en el procesamiento del lenguaje natural y la recuperación de información. Además, los aspectos presentados en forma de preguntas y respuestas fueron identificados colaborativamente y muestran el interés de los estudiantes de la asignatura “Semántica y pragmática en la web” del Máster en Tecnologías de la Lengua de la UNED.

Como mis estudiantes han señalado, llama la atención tanto la utilización repetida de varias técnicas, como el hecho de que muchas veces gran parte del trabajo realizado por los grupos participantes consiste en adaptar a cada tarea los métodos, modelos y recursos más adecuados. Y en la mayor parte de los casos, parece que esta adaptación tiene mucho de refinamiento, prueba y medida de resultados hasta llegar a un resultado que el investigador toma como bueno.

Además, espero que al identificar sus contribuciones recuerden los comentarios en el foro sobre cómo y cuánto de importante es referenciar en los textos, para que el lector pueda complementar la información aportada (yo he necesitado 70 links en el texto y 55 artículos en la bibliografía).

Mi último comentario es un especial agradecimiento a mis estudiantes, que hicieron posible tanto esta compilación como mi aprendizaje.

Referencias

[1] *Aiala, Rosá; Irene Castellón*, *Luis Chiruzzo*, *Hortensia Curell*, *Mathias Etcheverry*, *Ana ernández*, *Gloria Vázquez*, *Dina Wonsever* (2019) Overview of FACT at IberLEF 2019: Factuality Analysis and Classification Task. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 105-110.

[2] *Baly, Ramy; Georgi Karadzhov, Dimitar Alexandrov, James Glass, Preslav Nakov* (2018) Predicting Factuality of Reporting and Bias of News Media Sources. https://arxiv.org/pdf/1810.01765.pdf

[3*] Barnes, J.* (2019) LTG-Oslo Hierarchical Multi-task Network: The Importance of Negation for Document-level Sentiment in Spanish. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421. PP: 378-389.

[4] *Beltran, J., Gonzalez, M.* (2019) Detection of Negation Cues in Spanish: The CLiC-Neg System. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421. PP: 352-360.

[5] *Benavent, Joan; X. Benavent, E. de Ves, R. Granados, A. García-Serrano*. (2010) Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches. M. Braschler, D. Harman, E. Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, ISBN 978-88-904810-0-0, 22-23 September 2010, Padua, Italy**.** CEUR Proceedings, Volume 1176. ISSN1613-0073.

[6*] Catalá, N., Martin, M.* (2019) Coin flipper at ehealth-kd challenge 2019: Voting lstms for key phrases and semantic relation identification applied to Spanish ehealth texts In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421. PP: 17-25.

[7] *Casavantes, M., López, R., González, L.C.* (2019) Uach at mex-a3t 2019: Preliminary results on detecting aggressive tweets by adding author information via an unsupervised strategy. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421. PP: 537-543.

[8] *Castro Castro, D., Artigas Herold, M.F., Ortega Bueno, R., Muñoz, R.*: Cerpamidua at MexA3T (2019) Transition point proposal. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421. PP: 502-507.

[9] *Chiruzzo*, *Luis; Santiago Castro*, *Mathias Etcheverry*, *Diego Garat*, *Juan José Prada*, *Aiala Rosá* (2019) Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421. PP: 132-144.

[10] *Collovini*, *Sandra; Joaquim Santos*, *Bernardo Consoli*, *Juliano Terra*, *Renata Vieira*, *Paulo Quaresma*, *Marlo Souza*, *Daniela Barreiro Claro*, *Rafael Glauber*. (2019) IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 390-410.

[11] *Cristóbal Colón-Ruiz, Isabel Segura-Bedmar.* (2019) Protected Health Information Recognition by BiLSTM-CRF. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 679-686.

[12] *Díaz-Galiano*, *Manuel Carlos; Manuel García-Vega*, *Edgar Casasola*, *Luis Chiruzzo*, *Miguel García-Cumbreras*, *Eugenio Martínez Cámara*, *Daniela Moctezuma*, *Arturo Montejo Ráez*, *Marco Antonio Sobrevilla Cabezudo*, *Eric Tellez*, *Mario Graff*, *Sabino Miranda*. (2019) Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 550-560.

[13] *Ezra Aragón*, *Mario; Miguel Álvarez-Carmona*, *Manuel Montes-y-Gómez*, *Hugo Jair Escalante*, *Luis Villaseñor-Pineda*, *Daniela Moctezuma*. (2019) Overview of MEX-A3T at IberLEF 2019: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 478-494.

[14*] Fabregat, H., Duque, A., Martinez-Romo, J., Araujo, L.:* (2019) Extending a Deep Learning approach for Negation Cues Detection in Spanish. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 663-670.

[15] *Ferro N. and C. Peters (Eds)* (2019) Information Retrieval Evaluation in a Changing World: Lessons learned from 20 Years of CLEF. INRE, V 41, Springer. https://doi.org/10.1007/978-3-030-22948-1

[16] *Fonseca, Evandro B; André Antonitsch, Sandra Collovini, Daniela Amaral,Renata Vieira, and Anny Figueira*. (2016) Summ-it++: an enriched version of the summ-it corpus. In Proceedings of the Tenth Int. Conf. on Language Resources and Evaluation (LREC'16), pages 2047-2051.

[17] *Gamallo, P., Garcia, M.* (2015) Multilingual open information extraction. In: Proceedings of Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence, EPIA 2015. pp. 711-722. Coimbra, Portugal.

[18] *Gamallo, P., Garcia, M.:* (2017) Linguakit: uma ferramenta multilingue para a analise linguistica e a extracao de informacao. Linguamatica 9(1), 19-28.

[19] *Giudice, V.* (2019) Aspie96 at FACT (IberLEF 2019): Factuality Classification in Spanish Texts with Character-Level Convolutional RNN and Tokenization. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 119-125.

[20] *Glauber, R., de Oliveira, L.S., Sena, C.F.L., Claro, D.B., Souza, M.* (2018) Challenges of an annotation task for open information extraction in portuguese. In: International Conference on Computational Processing of the Portuguese Language. pp. 66-76. Springer.

[21] *Godino, I.G., DHaro, L.F.* (2019) Gth-upm at TASS 2019: Sentiment analysis of tweets for spanish variants. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 579-588.

[22] *Gonzalez, J. A., Hurtado, L.F., Pla, F.:* (2019) Elirf-upv at TASS 2019: Transformer encoders for twitter sentiment analysis in spanish. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 571-578.

[23] *Gonzalez, J.A.; Hurtado, L.F.; Pla, F.:* (2019) ELiRF-UPV at IroSvA: Transformer Encoders for Spanish Irony Detection. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 278-284.

[24] *Hassan, Fadi; Mohammed Jabreel; Najlaa Maarrof; David Sánchez; Josep Domingo-Ferrer; Antonio Moreno.* (2019) ReCRF: Spanish Medical Document Anonymization using Automatically-crafted Rules and CRF. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 727-734.

[25] *Ismailov, A.:* Humor Analysis Based on Human Annotation Challenge at IberLEF 2019: First-place Solution. (2019) In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 160-164.

[26] *Jabreel*, *Mohammed; Fadi Hassan*, *Najlaa Maarrof*, *David Sánchez*, *Josep Domingo-Ferrer*, *Antonio Moreno* (2019) E2EJ: Anonymization of Spanish Medical Records using End-to-End Joint Neural Networks. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 712-719.

[27] *Jiménez-Zafra*, *Salud María; Noa Patricia Cruz Díaz*, *Roser Morante*, *María-Teresa Martín-Valdivia*. (2019) NEGES 2019 Task: Negation in Spanish. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 329-341.

[28] *Jiang, Dehuan; Yedan Shen; Shuai Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Ruifeng Xu, Jun Yan, Yi Zhou.* (2019) A Deep Learning-Based System for the MEDDOCAN Task. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 761-767.

[29] *Lastra J.J.; A. Garcia-Serrano* (2015) A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. International Scientific Journal Engineering Applications of Artificial Intelligence. V 46 PP: 140-153, PERGAMON-ELSEVIER SCIENCE.
http://dx.doi.org/10.1016/j.engappai.2015.09.006

[30] *Lastra-Díaz, J.J; Josu Goikoetxea; Mohamed Ali Hadj Taieb; Ana García-Serrano; Mohamed Ben Aouicha; Eneko Agirre*. (2019) A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. Engineering Applications of Artificial Intelligence, Volume 85, Pages 645-665. https://doi.org/10.1016/j.dib.2019.104432

[31] *Lara-Clares, Alicia; Garcia-Serrano, Ana;* (2019) Key Phrases Annotation in Medical Documents: MEDDOCAN 2019 Anonymization Task. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 755-760. https://pdfs.semanticscholar.org/0b08/3d574f5248cad493cd157e036a7f67df2ef8.pdf

[32] *Lara-Clares, Alicia; Garcia-Serrano, Ana;* (2019) LSI2_UNED at eHealth-KD Challenge 2019: A Few-shot Learning Model for Knowledge Discovery from eHealth Documents. In Proceedings of IberLEF 2019

(Bilbao). CEUR-WS Vol 2421, PP: 60-66.
https://pdfs.semanticscholar.org/4f8c/3b91aed7bae2f2a84d34d830f93a77d991dd.pdf

[33] *Lopez-Ramos y Arco-Garcia L.*, (2019) Aprendizaje profundo para la extracción de aspectos en opiniones textuales. Revista cubana de ciencias informáticas, V13, N2.
http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992019000200105&lng=es&nrm=iso&tlng=es

[34] *Lukas Lange, Heike Adel, Jannik Strötgen.* (2019) NLNDE: The Neither-Language-Nor-Domain-Experts' Way of Spanish Medical Document De-Identification. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 671-678.

[35] *Luque, F.M.* (2019) Atalaya at tass: Data augmentation and robust embeddings for sentiment analysis. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 561-570.

[36] *Marimon*, *Montserrat; Aitor Gonzalez-Agirre*, *Ander Intxaurrondo*, *Heidy Rodríguez*, *Jose Lopez Martin*, *Marta Villegas*, *Martin Krallinger. (2019)* Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 618-638.

[37] *Martínez-Fernández, Paloma; Garcia Serrano, Ana* (2002) Utilizando recursos lingüísticos para mejora de la recuperación de información en la Web Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 6, núm. 16, verano 2002, pp. 55-64.

[38] *Martínez-Fernández, J.L.; Julio Villena Román; A. M. García-Serrano; José Carlos González-Cristóbal* (2005) Combining Textual and Visual Features for Image Retrieval Accessing Multilingual Information Repositories Lecture Notes in Computer Science V 4022 PÁGINAS: 80-691 Springer-Verlag Berlin.

[39] *Medina, S., Turmo, J.* (2019) Talp-upc at ehealth-kd challenge 2019: A joint model with contextual embeddings for clinical information extraction. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 78-84.

[40] *Miranda-Belmonte, H.U., Lopez-Monroy, A.P.*: Early Fusion of Traditional and Deep Features for Irony Detection in Twitter (2019) In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 272-277.

[41] *Ortega-Bueno*, *Reynier; Francisco Rangel*, *Delia Irazú Hernández Farías*, *Paolo Rosso*, *Manuel Montes-y-Gómez*, *José Medina-Pagola*. (2019) Overview of the Task on Irony Detection in Spanish Variants. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 229-256. Bilbao.

[42] *Pablo Sánchez, César de; Martínez Fernández, José Luis; Martínez Fernández, Paloma; Villena Román, Julio; García Serrano, Ana; Goñi Menoyo, José Miguel y González Cristóbal, José Carlos* (2004). miraQA: Initial experiments in Question Answering. Proc. "5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004", 15/09/2004-17/09/2004, Bath, Reino Unido.
http://oa.upm.es/4697/

[43] *Piad-Morffis*, *Alejandro; Yoan Gutiérrez*, *Juan Pablo Consuegra-Ayala*, *Suilan Estevez-Velarde*, *Yudivián Almeida-Cruz*, *Rafael Muñoz*, *Andrés Montoyo* (2019) Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 1-16.

[44] *Premjith, B., Soman, K.P., Poornachandran, P.* (2019) Amrita CEN@FACT: Factuality Identification in Spanish Text. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 111-118.

[45] *Saurí, R.* (2008) A Factuality Profiler for Eventualities in Text. Brandeis University.

[46] *Saurí, R., Pustejovsky, J.* (2009) Factbank: a corpus annotated with event factuality. Language resources and evaluation 43(3), 227.

[47] *Sánchez-León, Fernando.* (2019) Resource-Based Anonymization for Spanish Clinical Cases. In Proceedings of IberLEF 2019 (Bilbao). CEUR-WS Vol 2421, PP: 704-711.

[48] *Sena, C.F.L., Claro, D.B*. (2018) Pragmatic information extraction in brazilian Portuguese documents. (2018) In: International Conference on Computational Processing of the Portuguese Language. pp. 46-56. Springer.

[49] *Sena, C.F.L., Claro, D.B.* (2019) Inferportoie: A portuguese open information extraction system with inferences. Natural Language Engineering 25(2), 287-306

[50] *Sena, C.F.L., Glauber, R., Claro, D.B.* (2017) Inference approach to enhance a Portuguese open information extraction. In: ICEIS (1). PP. 442-451.

[51] *Smith, N.A.* (2019) Contextual Word Representations: A Contextual Introduction. arXiv:1902.06006v2.

[52] *Sánchez-Rada, J. Fernando; Carlos Angel Iglesias.* (2019). Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. Information Fusion, 52, PP: 344-356.

[53] *Suárez-Paniagua*, V.; *Isabel Segura-Bedmar*; *Paloma Martínez* (2018) LABDA at TASS-2018 Task 3: Convolutional Neural Networks for Relation Classification in Spanish eHealth documents, Ceur-WS Vol 2172. PP:71-76.

[54] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. Ukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008. Curran Associates, Inc.