DeepText

Nueva generación de modelos neuronales de inteligencia artificial para transformar las tecnologías de la lengua en la industria del País Vasco

Entregable E3.1+E3.2+E3.3: Resultados de adaptación a dominio y a idioma de los modelos de lenguaje

Responsable	**Vicomtech**
Tipo	**Informe**
Ejercicio	**2021**

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E3.1+E3.2+E3.3

Índice

ELKARTEK 2020
Proyectos de investigación fundamental colaborativa

DeepText – Entregable E3.1+E3.2+E3.3

1. Introducción

Este entregable aglutina información sobre los resultados de las diferentes tareas relacionadas con el PT3 del proyecto DeepText. En este documento se combinan los entregables E3.1, E3.2 y E3.3 debido a la alta interrelación de las tareas y resultados obtenidos. La adaptación a dominio y a idioma van generalmente de la mano al utilizar modelos de lenguaje basados en Deep Learning, ya que todo pivota generalmente en torno a los datos, la disponibilidad de datos adecuados, y al uso y explotación que se hace de esos datos.

2. Feature-based vs. Fine-tuning

Los modelos de Deep Learning basados en Transformers basan sus facultades en la capacidad que tienen para emitir representaciones vectoriales contextuales de los tokens (palabras y sub-palabras) que reciben a la entrada.

Estas representaciones a su vez pueden explotarse de diferentes maneras para proyectarlas a un espacio de categorías asociadas a la resolución de un problema concreto.

A las representaciones vectoriales obtenidas mediante modelos de Transformers (es decir, mediante su mecanismo interno de self-attention) se les denomina contextual-word-embeddings. En comparación a sus equivalentes estáticas (word2vec, GloVE, etc.), a cada palabra posible del vocabulario que se haya modelado no le corresponde una única representación vectorial (word-embedding) sino que para cada contexto en el que dicha palabra aparezca, su embedding se adapta en consecuencia. Esto es fundamental, ya que permite un modelado semántico mucho más flexible y preciso, al tener en cuenta todo el texto para representar una palabra, en lugar de la palabra descontextualizada.

Para conseguir que un modelo de Transformers emita embeddings contextuales para un idioma/dominio dados, hay que pre-entrenarlo. Como cualquier otro modelo matemático, sin un entrenamiento previo que ajuste las variables internas del mismo (los parámetros del modelo), los embeddings que se emitidos serían meros números aleatorios sin significado ni utilidad. El pre-training de un modelo de Transformers se realiza mediante tareas denominadas “self-supervised”. El nombre deviene del hecho de que, aunque se trata de una tarea de aprendizaje supervisado, no se necesita etiquetar manualmente el texto con el que el modelo va a entrenar, ya que se aprovecha la propia consistencia y coherencia del lenguaje natural para que el modelo aprenda a interpretarlo. La tarea de aprendizaje “self-supervised” más común es el Masked Language Modelling, que consiste en enmascarar un porcentaje determinado de palabras, y forzar al modelo a que intente predecir los huecos en base al contexto. Otro tipo de tareas consisten en añadir ruido, permutar palabras,

eliminar fragmentos o pedir al modelo que indique si dos frases dadas son continuación una de la otra.

Este tipo de tareas se pueden derivar de manera automática a partir de grandes volúmenes de textos. Cuando el modelo a aprendido a aproximar la tarea en cientos o miles de millones de textos, los embeddings que emite empiezan a seguir patrones y a portar información lingüística explotable.

El pre-training, en la medida que requiere un volumen enorme de datos para permitir al modelo ir aprendiendo gradualmente a modelar el idioma (o idiomas), es una tarea computacionalmente muy costosa, y se suele realizar una única vez, para luego reutilizar el modelo base entrenado en diferentes tareas.

El entrenamiento de un modelo de lenguaje pre-entrenado para una nueva tarea específica es lo que se conoce como fine-tuning.

En general se trata de permitir que el modelo base emita los word-embeddings contextuales correspondientes, y usar estos a su vez como entradas de más capas neuronales que finalmente emiten la respuesta adecuada (en función del problema a resolver).

El fine-tuning es un proceso más ligero que el pre-training, porque aunque también es recomendable un volumen determinado de datos de entrenamiento, el orden de magnitud es muy inferior. Dependiendo de la dificultad de la tarea, con unas pocas decenas o centenares de ejemplos de cada clase a predecir puede ser más que suficiente. Aún así, en la medida que los modelos de Transformers son pesados, incluso en sus variantes más ligeras, actualizar todos los pesos de un modelo de Transformer es algo que conlleva una carga computacional nada desdeñable, y que requiere del uso de hardware especializado (GPUs) para poder realizarlo en un tiempo razonable.

Una alternativa es no actualizar todos los pesos de un modelo de Transformers, sino dejar los pesos que venía pre-entrenados, congelados sin actualizar, y hacer el “fine-tuning” sólo sobre una última serie de capas neuronales específicas para la tarea en cuestión. Al tratarse de capas mucho más ligeras, el entrenamiento es mucho más rápido, a costa de perder cierta capacidad de aprendizaje al estar gran parte del modelo “congelado”. Dado que esta modalidad es equivalente a usar el modelo de Transformer para convertir el texto en vectores de características (feature-vectors), se puede denominar “feature-based”.

Para comprobar como afecta el trade-off entre facilidad/velocidad de entrenamiento y capacidad de aprendizaje entre un full-fine-tuning (entrenamiento del modelo completo) y un feature-based training (sólo entrenar las capas finales), hemos hecho una serie de pruebas.

En concreto hemos realizado una serie de entrenamientos de modelos basados en la tarea de Named Entity Recognition and Classification (NERC). Esta tarea consiste en la detección de entidades nombradas (personas, lugares, organizaciones, etc.) en textos. La tarea es lo de menos, ya que lo que queremos comparar en las dos variantes de entrenamiento en igualdad de condiciones. Se han

hecho las pruebas tanto para castellano como para euskera, utilizando diferentes modelos base. En concreto:

- Para castellano se ha utilizado el dataset CoNLL2022 de entidades nombradas, y los modelos base BETO (BERT para castellano), IXAmBERT (BERT para castellano, euskera e inglés) y mBERT (BERT multilingüe para 104 idiomas). Se han hecho pruebas tanto de “fine-tuning” como “feature-based”.
- Para euskera se ha utilizado el dataset Egunkaria de entidades nombradas, y los modelos base BERTeus (BERT para euskera), IXAmBERT (BERT para castellano, euskera e inglés), y mBERT (BERT multilingüe para 104 idiomas). Se han hecho pruebas tanto de “fine-tuning” como de “feature-based”.

Las siguientes imágenes muestran los resultados de las pruebas de manera gráfica. Las gráficas muestran las curvas de las métricas de validación, que indican, a lo largo del entrenamiento, cómo de bien está funcionando el modelo en un conjunto de datos reservado a tal efecto.

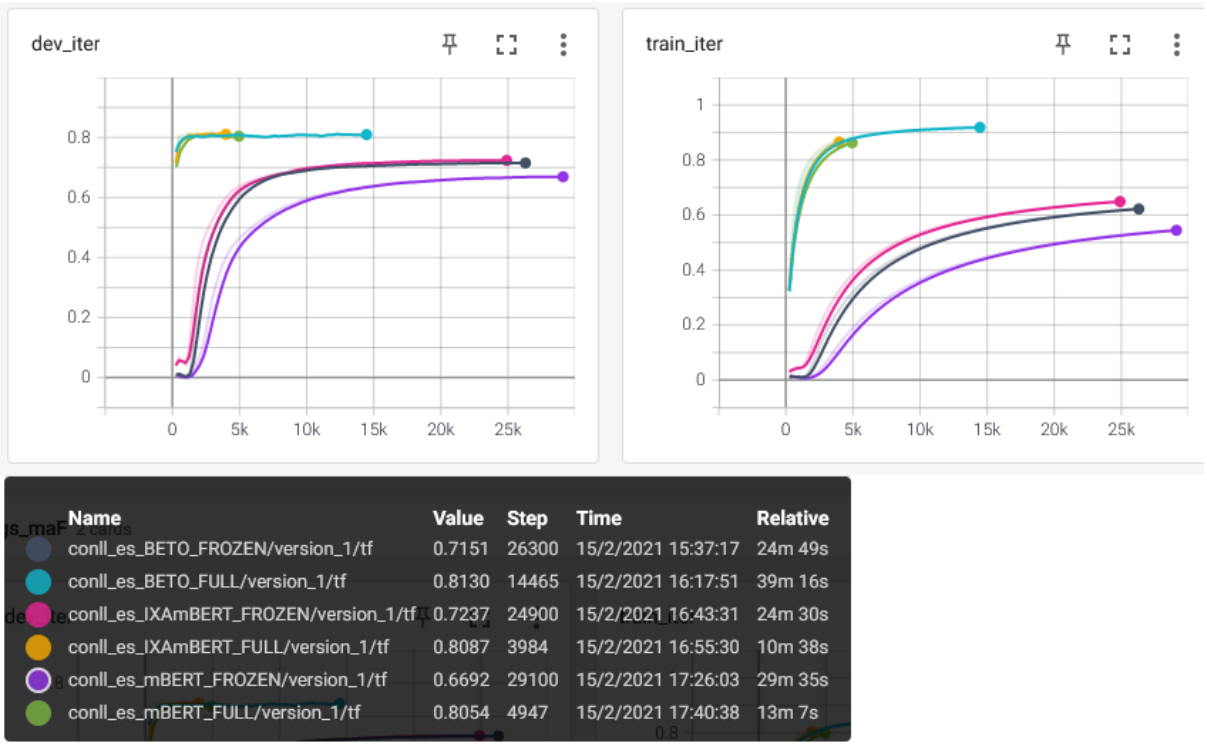

Figura 1. Comparación "feature-based" vs. "full fine-tuning" en castellano

Figura 2. Comparación "feature-based" vs. "full-fine-tuning" en euskera

La métrica empleada es la F1 micro-ponderada, habitual en este tipo de problemas. Esta métrica es una media armónica entre la precisión y la cobertura de las predicciones realizadas por el modelo. Cuanto más alta sea, mejor está efectuando el modelo su labor.

Las gráficas van acompañadas de una leyenda en la que se indica el código de color de las líneas y su correspondencia con los diferentes experimentos realizados. Aquellos experimentos que llevan “FROZEN” en el nombre, son los que tienen “congelados” los modelos de Transformer, y por lo tanto son “feature-based” (es decir, el modelo de Transformer no se entrena durante el fine-tuning, sino que se usa tal cual, para generar representaciones, features, de las palabras). Los experimentos que llevan “FULL” en el nombre son aquellos en los que se ha realizado un full-fine-tuning, es decir, que el modelo completo, incluyendo las capas de Transformer, han sido entrenadas.

Es fácil observar la diferencia entre ambas variantes en todos los experimentos. Las versiones “feature-based” son mucho más rápidas de entrenar. Esto se puede observar en el “step” (número de pasos de entrenamiento) por el tiempo utilizado (columna “relative”). Mientras que los modelos “FULL” realizan entre 4k y 5k “steps” en el orden de 10 minutos, los “FROZEN” hacen el doble o el triple. No obstante, los modelos “FROZEN” llegan a cierto techo del que no pueden pasar, debido a

que disponen de menos flexibilidad para aprender. Los únicos pesos que pueden adecuar para resolver la tarea son los de una capa de clasificación al final, mientras el resto de los parámetros del modelo de Transformers permanece congelado. Los modelos “FULL” en cambio, aunque requieren de mayor esfuerzo computacional, pueden desplegar todo su poder de expresión a la hora de acomodar los datos del problema, y, como se ve en las gráficas su techo queda muy por encima.

Esta diferencia tiende a acrecentarse cuanto más complejo es el problema. Cuando el problema es relativamente sencillo, un modelo sencillo puede dar una respuesta suficiente. Cuando el problema es complejo, y hay muchos datos de entrenamiento, la cantidad de parámetros del modelo que intervienen en el aprendizaje cobra mayor relevancia.

En conclusión, en la medida de lo posible, y atendiendo a la complejidad del problema y a la disponibilidad de datos de entrenamiento, es recomendable entrenar el modelo completo, incluyendo todas las capas de Transformers. Obviamente, esto conlleva un coste computacional elevado, que habrá que evaluar en cada caso, pero el resultado será un modelo entrenado mucho más capaz de adecuarse al problema.

3. Transferencia entre dominios e idiomas

Esta sección engloba los diversos trabajos realizados en relación al uso de modelos de Transformer adaptados a diferentes dominios e idiomas. Como las diferentes tareas de procesamiento de texto se enmarcan siempre para un dominio en concreto, y por supuesto, en un idioma dado, los trabajos abordan ambas facetas de manera simultánea de modo inherente.

En general, la adaptación o transferencia de conocimiento en relación al idioma está más basada en el uso de modelos de lenguaje, con arquitectura de Transformers, pre-entrenados para los idiomas requeridos. En este caso se hace un uso intensivo tanto del modelo BERTeus (modelo BERT exclusivo para euskera entrenado por el grupo IXA), y el modelo IXAmBERT (modelo BERT trilingüe euskera-castellano-inglés). Por supuesto, a la hora de hacer “fine-tuning” para enseñar al modelo a resolver una tarea específica, la disponibilidad de datos de entrenamiento en el idioma objetivo es también un factor de gran importancia.

En algunos casos, por ejemplo, clasificación de documentos, la “transferencia de conocimiento”, o “transfer-learning” pueden conseguir lo que se denomina “zero-shot transfer-learning”. Esto significa que un modelo entrenado para una tarea, con datos de un idioma específico, es capaz de resolver (con cierto grado de eficacia) la misma tarea en un segundo idioma para el cuál no se han dispuesto datos de entrenamiento. Por supuesto la efectividad del modelo suele ser inferior con respecto al idioma para el que sí se ha entrenado, y si la tarea es muy compleja generalmente la efectividad se degradará hasta puntos donde no es practicable usar dicho modelo. Lo ideal siempre es disponer de datos propios del idioma, aún si son limitados o generados de manera semi-automática.

A continuación, se listan las tareas donde se han aplicado estos modelos, para resolver diferentes problemas, en diferentes dominios de aplicación (medios de comunicación, salud, etc.) y en diferentes idiomas (euskera, castellano, inglés). La mayor parte de las tareas descritas han sido respaldadas por sus respectivos artículos científicos publicados en diversas conferencias. Por ello la descripción que se da en este documento es de corte informativo/divulgativo, sin entrar del todo en

algunos detalles, que en caso de interés, pueden consultarse en dichos artículos (citados al final del documento).

3.1 Clasificación de noticias en euskera

La clasificación de documentos es la tarea que consiste en asignar a cada texto una categoría, de un inventario predefinido de categorías que aplican al problema. En el caso de clasificación supervisada, el modelo en cuestión aprende a partir de ejemplos previos, en los que ya se ha indicado la categoría adecuada a una serie de documentos. El modelo aprende de estos ejemplos para obtener la capacidad de asignar de manera autónoma una etiqueta a nuevos documentos que no haya visto antes.

Partiendo de un conjunto de datos de noticias en euskera de un grupo de comunicación de Euskadi, se ha llevado a cabo el consiguiente entrenamiento de un clasificador. El espacio de categorías era de tamaño 10, incluyendo: **ekonomia, euskara, gizartea, hezkuntza, ingurumena, iritzia, jendeartea, kirola, cultura, politika**.

El clasificador está basado en un modelo de Transformers, en concreto el modelo BERT. Inicialmente se optó por probar el modelo BERT multilingüe, ya que es un modelo versátil que cuenta con conocimiento pre-entrenado para 104 idiomas, incluyendo el euskera. No obstante, debido a la amalgama de idioma que aglutina este modelo, tanto la representación de las palabras (espacio de tokens del tokenizador asociado al modelo), como el conocimiento de euskera del que dispone, son subóptimos con respecto a un modelo totalmente especializado para euskera.

Por ello, acto seguido se optó por reemplazar el modelo de base por el BERTeus, modelo BERT entrenado específicamente para euskera. Con ello se obtuvo una mejora de varios puntos en las métricas de evaluación.

El resultado del clasificador alcanzada una Fscore de 76%. Es un valor razonablemente alto y que ofrece un rendimiento muy satisfactorio. El valor de la métrica, así como su interpretación y su comportamiento para cada una de las clases involucradas puede observarse en la siguiente matriz de confusión.

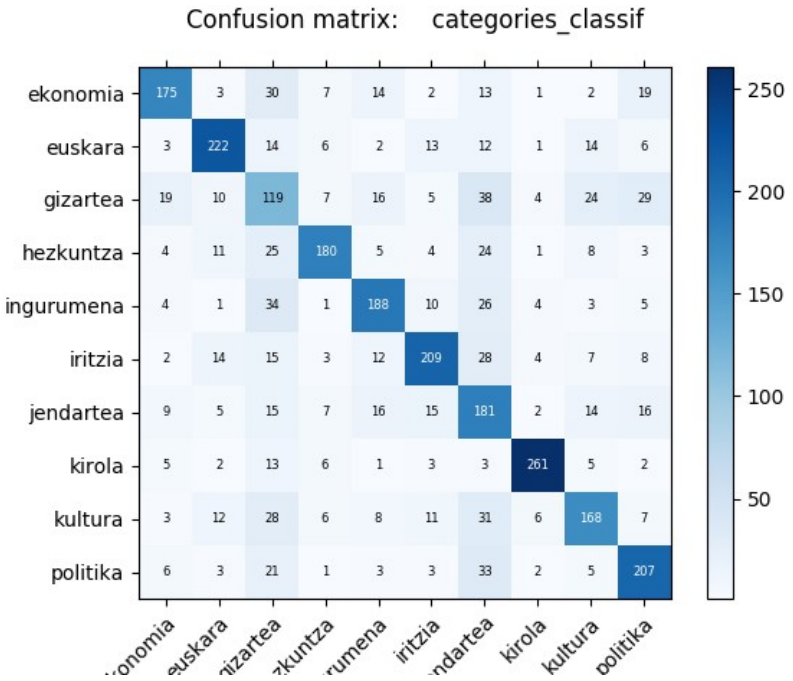

Figura 3. Matriz de confusión de clasificación de categorías en euskera

Como se observa en la matriz de confusión, la diagonal principal está muy marcada. Los elementos de la diagonal principal son aquellos en los que el modelo ha asignado la categoría correcta con respecto a la que se indicaba manualmente en el conjunto de datos de validación.

Si se observa detenidamente, además de ver que todas las categorías reciben un tratamiento muy adecuado por el modelo, se puede ver cuál es la mayor fuente de confusión. Las categorías que semánticamente son más ambiguas o próximas entre sí son las que más errores inducen al modelo, en este caso “gizartea” y “jendeartea”. Esto es normal, si tenemos en cuenta que son categorías muy amplias, que puede cubrir en horizontal otro espectro de noticias y entremezclarse con noticias que a su vez pertenecen a categorías diferentes. La propia definición de las categorías “gizartea” y “jendeartea” podría llevar a confusión a un humano, y de hecho es muy posible que en el conjunto de datos de entrenamiento no haya un criterio claro de cuándo una noticia ha recibido una categoría o la otra. También hay casos donde aplicarían varias categorías (por ejemplo una noticia sobre el euskera en el sistema educativo podría ser de “euskara” o de “hezkuntza”, y una noticia sobre la visita del lehendakari a unas ikastolas podría tratarse como “hezkuntza” o como “politika”).

En cualquier caso, más allá de la problemática concreta de este conjunto de datos, el comportamiento del modelo de clasificación para euskera es notable, y puede usarse directamente para dar soporte a la categorización automática de noticias en un entorno de producción real.

3.2 Resumen abstractivo en euskera

El resumen abstractivo consiste en partir de un texto de una cierta longitud, y sintetizar un nuevo texto, más breve, que recoja los aspectos e ideas principales del texto original. A diferencia del resumen "extractivo", en el que se seleccionan fragmento (generalmente frases) relevantes del texto original para componer el texto de salida, el resumen abstractivo escribe desde cero un texto completamente nuevo a partir del texto original. En este sentido es una tarea mucho más desafiante, ya que el modelo tiene que saber "leer" y "escribir" en el idioma objetivo, además de ser competente en su tarea de "resumir".

El resumen abstractivo es una tarea de procesamiento de lenguaje natural que se haya en TRLs más bajos. Incluso para inglés sigue siendo una labor compleja y cuyos resultados todavía admiten margen de mejora. Para euskera, aunque el estado del arte está avanzando muy deprisa, es aún más complejo debido por un lado a la falta de recursos específicos, y por otro a las complejidades intrínsecas adicionales que plantea un idioma aglutinante.

No obstante, se ha hecho una labor prospectiva para examinar qué tal funciona, y cuál sería el resultado esperado a corto y medio plazo de una tecnología semejante para euskera.

El primer escollo que sortear, tanto para euskera como para otros idiomas a la hora de abordar tareas de resumen automático, es la obtención de datos. En general la generación (manual, por humanos) de resúmenes es una tarea que requiere una alta carga cognitiva (más que indicar una etiqueta para un texto), además de ser una tarea con una alta carga de subjetividad (dado un texto no todo el mundo otorgará la misma relevancia a cada pieza de información). Por ello no es habitual disponer de conjunto de datos de entrenamiento para resumen automático. Ni siquiera para inglés, en los que se suele usar lo que se conoce "supervisión distante". La supervisión distante (o distant-supervision en inglés) consiste en aprovechar información que ya se está generado o se ha generado de manera natural, para aproximar el resultado que se quiere que el modelo aprenda. En este caso, se suele aprovechar la entradilla de una noticia, o los highlights destacados por el periodista, como aproximación al resumen del cuerpo entero de la noticia. Esta aproximación no es del todo exacta, pero cumple su cometido.

Siguiendo esta misma metodología se han utilizado las entradillas de las noticias en euskera como resúmenes de su cuerpo principal. Para acotar y filtrar un poco aquellos ejemplos resultantes que pudieran no ser correctos, se han aplicado filtros para eliminar noticias en las que la proporción entre el cuerpo de la noticia y su entradilla no estuviese dentro de determinado rango.

Una vez recopilado el corpus se ha procedido a entrenar y hacer pruebas con varios modelos encoder-decoder. Los modelos encoder-decoder son modelos que constan de una parte que se encarga de procesar el texto de entrada (codificarlo, de ahí el nombre encoder), y otra parte que se encarga de generar la salida a partir de la información del encoder, el decoder.

Las métricas obtenidas con diferentes modelos se pueden observar en la siguiente figura:

Tabla 3: Resultados de los modelos de resumen automático sobre la partición de *test*

Model	Rouge-1	Rouge-2	Rouge-L
IXAmBERT	27.33	11.92	22.64
RoBasquERTa	22.12	8.76	17.69
mT5-small	19.69	7.49	15.66

Figura 4. Resultados evaluación de modelos entrenados para resumen automático en euskera

A continuación, se muestra un pequeño ejemplo de salida de resumen automático en euskera. El resultado aún requiere de mejora antes de ser una tecnología fiable y utilizable en un entorno de producción, pero la tecnología avanza muy rápido.

TEXTO ORIGINAL

Orain baino bisitari gehiago erakartzeko bi egitasmo daude Nerbioiko ur-jauziaren gainean. Arabako Foru Aldundiak 6,5 kilometroko ibilbide berdea egokituko du, Delika herria eta Nerbioiko ur-jauzia lotzeko, eta gainera, behatoki berri bat eta egurrezko zubi tibetar bat eraikiko ditu. Gaztela eta Leongo Junta ere Monte Santiagoko begiratokia ordezkatzeko proiektua aztertzen ari da, turista gehiago jasotzeko amoz.
Juntaren zifra ofizialen arabera, urtero, 20.000tik gora pertsona sartzen dira Nerbioiko ur-jauzia bisitatzera Monte Santiago parketik, eta 100.000 auto inguru jasotzen dituzte egokitutako aparkalekuetan.

RESUMEN OBTENIDO AUTOMÁTICAMENTE (MODELO ENTRENADO EN EUSKERA)

Arabako Foru Aldundiak urteko ibilbide berdea egokituko du, Delika herrian eta Nerbioiko ur-jauziaren gainean. Urtero bezala, 20.000tik gora pertsona sartzen dira Monte Santiago parketik, eta 100.000 perts [...]

Figura 5. Ejemplo de resumen automático generado para euskera

3.3 Detección de conceptos y relaciones en texto clínico en castellano (e inglés)

Para seguir explorando las capacidades de estos modelos se ha participado también en algunas competiciones internacionales de la comunidad científica. Este tipo de competiciones son muy útiles para comprobar la eficacia de algunas ideas y modelos, ya que proporcionan un conjunto de datos y un problema a resolver, que son un excelente banco de pruebas.

En este caso la competición eHealth-KD-2020, y la siguiente edición eHealth-2021, han servicio como banco de pruebas para probar la adaptación de este tipo de modelos a diferentes dominios (textos de ámbito clínico, noticias relacionadas con salud, etc.).

La tarea consiste en detectar conceptos y relacionarlos entre sí.

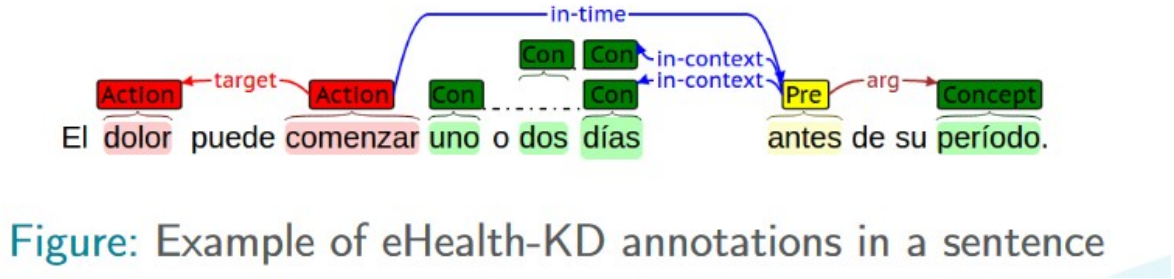

Figura 6. Ejemplo de deteción de conceptos y relaciones en texto del ámbito de la salud

El modelo utilizado para abordar la tarea ha sido tanto el BERT-multilingüe como el BETO (BERT específico para castellano). El uso de BETO proporcionaba de manera consistente varios puntos de ventaja sobre el BERT multilingüe, volviendo a demostrar la utilidad de disponer de un modelo adecuado adaptado al idioma. En la siguiente edición de la competición se utilizó el modelo base IXAmBERT para tener en cuenta el hecho de que los datos incluían contenido en inglés (para el cual no había datos de entrenamiento). El modelo de base IXAmBERT dispone de conocimiento de inglés y su espacio de tokens también estará más adecuado al procesamiento de contenido en inglés, lo cual le da una clara ventaja. Efectivamente, como se ve en la tabla de resultados, las variantes que usan IXAmBERT en este caso superan a los modelos que se basan en BETO (de sólo español).

En la siguiente imagen se muestra la tabla con los resultados de la participación en la competición:

		ES			EN			Total		
		P	R	F1	P	R	F1	P	R	F1
Main	Run 1: mT5	67.01	51.64	58.33	33.69	32.12	32.88	51.27	43.44	47.03
	Run 2: BETO	66.36	60.89	63.51	**37.12**	34.09	35.54	54.07	49.63	51.76
	Run 3: IXAmBERT *(best)*	**68.55**	**62.68**	**65.49**	35.41	**40.73**	**37.88**	52.75	**53.46**	**53.11**
	PUCs-UFMG *(2nd)*	-	-	-	-	-	-	**56.85**	50.28	52.84
A	Run 1: mT5	**83.46**	71.06	76.77	**56.33**	44.11	49.48	69.99	57.11	62.90
	Run 2: BETO	79.44	81.37	80.39	41.61	53.82	46.94	57.67	67.11	62.04
	Run 3: IXAmBERT *(2nd)*	79.60	**83.48**	**81.49**	50.79	**66.53**	**57.60**	63.10	**74.71**	68.41
	PUCs-UFMG *(best)*	-	-	-	-	-	-	**71.49**	69.73	**70.60**
B	Run 2: BETO	56.11	37.31	44.82	15.38	1.40	2.56	50.83	18.59	27.22
	Run 3: IXAmBERT *(2nd)*	**57.88**	**40.10**	**47.38**	**47.77**	**17.48**	**25.60**	**54.19**	28.31	37.19
	IXA *(best)*	-	-	-	-	-	-	45.36	**40.95**	**43.04**

Figura 7. Resultados de la competición sobre detección de conceptos y relaciones

La participación en la competición, además de arrojar conclusiones interesantes, concluyó con que el modelo propuesto fue el que obtuvo la mejor puntuación, alzándose como ganador de la competición, por dos años consecutivos.

3.4 Puntuación/capitalización automática de salida ASR en castellano y euskera

La puntuación y capitalización de salida de ASR consiste en añadir símbolos de puntuación y las mayúsculas pertinentes al texto transcrito por un sistema ASR (Automatic Speech Recognition). LA salida de un sistema de ASR, en general consiste en una cadena de palabras según hayan sido reconocidas por el sistema a partir del audio de entrada. No obstante, estas palabras suelen ser su versión minusculizada, y no figuran los símbolos de puntuación. Esto dificulta mucho la lectura y la posterior utilización de los textos transcritos resultantes, y por consiguiente es de gran interés poder restituir de manera coherente tantos las mayúsculas como los símbolos de puntuación.

Se ha llevado a cabo un conjunto de experimentos para entrenar un modelo que realice estas tareas de manera conjunta. Se ha realizado el entrenamiento tanto para procesar sólo textos en euskera, en castellano, o bilingües (capaz de procesar tanto texto en euskera como en castellano de manera simultánea).

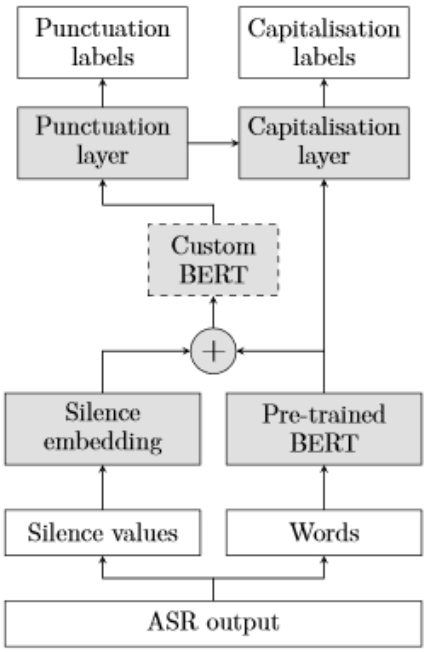

Figure 3: Main architecture of the BERT-based AUTOPUNCT system.

Figura 8. Diagrama simplificado de la arquitectura del modelo puntuación/capitalización

El modelo en cuestión además de tener en cuenta el texto (palabras) transcrito por el sistema ASR, también tiene en cuenta la duración de los silencios entre las palabras locutadas. De este modo se incorpora al modelado de información la intuición de que las diferentes pausas realizadas al hablar guardan correlación con los diferentes signos de puntuación del discurso.

La imagen muestra un diagrama simplificado de la arquitectura del modelo donde se combinan los modelos de Transformers con otras capas de redes neuronales para incluir la información de los silencios, y al mismo tiempo obtener tanto la salida de capitalización como la de puntación.

Los resultados se han comparado con sistemas anteriores que utilizan tecnologías anteriores a modelos de Transformers. En la tabla se observa el desglose de las diferentes variantes de modelos entrenados y evaluados, tanto para euskera, castellano y bilingüe.

La tabla muestra los valores de las métricas de evaluación para los diferentes signos de puntuación (PER=PERIOD, COM=COMMA, etc.).

Basque (EU)

S	D	*f*	B	COM	PER	QUES	COL	DASH	0_QS	SCOL	0_QUO	QUO	ELL	EXC	0_EX
×	–	–	×	74.6	84.8	58.8	**53.8**	13.6	0.0	**17.9**	59.3	47.6	0.0	0.0	0.0
×	–	–	✓	**75.7**	85.5	60.4	50.9	**20.3**	0.0	15.7	55.3	**54.5**	8.3	0.0	0.0
✓	×	–	×	75.3	85.8	61.0	51.2	0.0	0.0	0.0	54.4	40.0	0.0	0.0	0.0
✓	×	–	✓	75.5	85.0	63.5	47.1	3.6	0.0	0.0	50.7	43.9	0.0	0.0	0.0
✓	✓	G	×	75.4	85.6	**65.5**	49.8	14.5	0.0	10.9	58.1	51.2	0.0	0.0	0.0
✓	✓	G	✓	75.2	85.9	60.8	48.5	14.5	0.0	12.9	55.1	51.5	**11.6**	0.0	0.0
✓	✓	U	×	75.2	85.5	64.5	50.6	5.4	0.0	6.0	56.4	46.7	0.0	0.0	0.0
✓	✓	U	✓	75.5	86.1	62.0	50.0	13.6	0.0	10.3	**59.9**	47.3	0.0	0.0	0.0
BRNN				70.1	**87.2**	52.2	23.4	0.0	0.0	12.4	0.0	0.0	0.0	0.0	0.0

Spanish (ES)

S	D	*f*	B	COM	PER	QUES	COL	DASH	0_QS	SCOL	0_QUO	QUO	ELL	EXC	0_EX
×	–	–	×	79.8	85.7	58.5	49.8	17.6	64.2	30.6	48.9	26.9	5.4	5.6	10.7
×	–	–	✓	80.1	86.1	56.5	48.1	**21.7**	63.8	33.3	50.4	32.7	**16.3**	**13.3**	**24.1**
✓	×	–	×	79.8	86.8	56.3	**51.6**	11.9	61.3	23.8	48.4	25.3	0.0	2.9	0.0
✓	×	–	✓	79.7	85.1	56.6	46.1	12.3	63.3	25.5	50.0	20.2	0.0	0.0	5.9
✓	✓	G	×	80.1	86.5	54.1	45.0	17.9	62.0	27.1	47.7	31.1	0.0	2.9	5.7
✓	✓	G	✓	**80.2**	87.0	**59.7**	46.8	20.4	**64.8**	31.3	51.9	33.0	0.0	2.9	16.0
✓	✓	U	×	**80.2**	86.3	57.0	48.4	16.9	63.4	30.5	48.2	29.4	1.8	2.9	5.7
✓	✓	U	✓	80.0	**87.3**	59.0	48.2	21.4	64.0	**33.9**	**52.2**	**34.2**	0.0	5.7	16.2
BRNN				65.4	85.6	46.0	11.9	1.8	23.0	1.7	13.0	0.0	5.2	11.9	0.0

Spanish+Basque (ES+EU)

S	D	*f*	B	COM	PER	QUES	COL	DASH	0_QS	SCOL	0_QUO	QUO	ELL	EXC	0_EX
×	–	–	×	75.6	83.7	**56.9**	48.6	14.7	57.5	19.3	53.6	41.7	13.2	14.9	16.7
×	–	–	✓	75.9	83.7	55.7	48.6	21.1	59.9	20.8	55.2	**44.8**	**23.1**	**24.6**	**33.0**
✓	×	–	×	75.9	84.6	54.1	**51.7**	10.9	57.0	18.3	54.7	37.7	12.7	11.4	10.8
✓	×	–	✓	75.3	83.7	**56.9**	50.2	21.6	**60.3**	20.3	54.1	43.8	20.1	15.7	21.3
✓	✓	G	×	76.1	84.5	55.4	48.6	7.6	57.1	19.8	52.4	36.4	10.4	9.3	13.9
✓	✓	G	✓	**76.5**	85.4	56.2	51.5	18.1	58.3	21.3	55.3	40.8	18.5	19.0	24.0
✓	✓	U	×	76.3	85.0	55.8	51.3	16.4	57.9	22.1	54.4	40.7	17.4	13.3	24.1
✓	✓	U	✓	76.2	85.6	56.2	50.2	**23.7**	58.8	**24.8**	**55.5**	41.7	19.0	21.4	32.7
BRNN				69.7	**87.4**	49.9	16.1	2.0	3.3	9.1	6.2	0.0	13.6	5.2	0.0

Table 5: Class-wise F_1 scores for AUTOPUNCT and the BRNN-based system in each language. Labels with a F_1 score of 0.0 in the three language scenarios were omitted. S: Using information of silences. D: Using discrete silences. *f*: Gaussian (G) or uniform (U) distribution for contiguous buckets. B: Adding a custom BERT.

Figura 9. Tabla de resultados evaluación de las diferentes variantes de modelo de puntuación/capitalización

3.5. Sustitución de entidades con información sensible en euskera y castellano

La anonimización de textos consiste en detectar y eliminar información sensible que pueda aparecer escrita en documentos, tales como nombres de personas, datos de contacto, etc.

Un proceso de anonimización se puede descomponer en dos pasos:

- la detección de las palabras que portan información sensible, así como su clasificación en distintos tipos, y
- la eliminación, por ofuscación o reemplazo, de la información sensible detectada

Se han realizado diferentes pruebas para desarrollar modelos que permitan llevar a cabo este tipo de tareas relacionadas con la anonimización, tanto en euskera como en castellano, de manera simultánea.

Tomando como base el modelo de lenguaje IXAmBERT (BERT entrenado por el grupo IXA para euskera, castellano e inglés), se han entrenado modelos de reconocimiento de entidades, y se ha integrado un sistema de sustitución neuronal para tratar de generar palabras de reemplazo que mantengan la coherencia contextual (concordancia de género, de tipo de entidad, terminaciones morfológicas, etc.).

La imagen muestra un ejemplo de un texto, donde se han coloreado las palabras detectadas como portadoras de información sensible (según el modelo entrenado, es decir, personas, lugares, etc.). En la parte inferior de la imagen aparece el mismo texto, tras haber aplicado la sustitución basada en un modelo neuronal. El resultado, como puede verse, es un documento similar, que puede leerse de manera natural y tratarse como un texto equivalente, pero en el que la información original considerada sensible ha sido eliminada.

Una ventaja adicional de reemplazar la información por otra equivalente en lugar de simplemente eliminar o reemplazarla por caracteres arbitrarios (‘xxxxx’, o ‘*****’), es que aunque el modelo se equivoque y falle en la detección de alguna entidad aislada, a la salida sería difícil percibirlo. A priori todas las entidades tienen un aspecto natural, de no haber sido anonimizadas, por lo que es difícil inferir que una entidad presente en la salida del proceso sea real o una entidad de reemplazo.

Figura 10. Ejemplo de salida de procesamiento para anonimización (detección y reemplazo de información sensible) en euskera y castellano

3.6. Sistemas de búsqueda de respuesta

El estado del arte de los sistemas de búsqueda de respuestas (QA por su abreviatura en inglés de *Question Answering*) viene marcado por el uso de modelos de lenguaje pre-entrenados que son ajustados para la tarea específica de QA. Posteriormente, estos modelos se pueden modificar según el objetivo del sistema de QA. Así en el proyecto DeepText hemos creado varios modelos de QA para diferentes dominios e idiomas utilizando diferentes conjuntos de datos para ajustar el modelo original de QA:

- Modelo de QA monolingüe para inglés de propósito general. Este modelo está basado en el modelo de lenguaje BERT y hemos utilizado el conjunto de preguntas y respuestas QuAC de dominio general para ajustar ese modelo.
- Modelo de QA monolingüe para inglés para los dominios de cocina, viajes y películas. Este modelo está basado en el modelo de lenguaje BERT y hemos utilizado el conjunto de preguntas y respuestas DoQA que contiene 3 subconjuntos de esos 3 dominios diferentes para ajustar ese modelo.
- Modelo de QA multilingüe (inglés, euskera y español) que es capaz de responder preguntas sobre una persona o algún otro artículo que está en Wikipedia. Este modelo está basado en el modelo de lenguaje IXAmBERT y hemos utilizado los conjuntos de preguntas y respuestas QuAC (inglés) y ElkarHizketak (euskera). Este sistema ha sido integrado en el asistente virtual de Google Assistant, por lo que se puede acceder desde teléfonos móviles o altavoces inteligentes Google Home (https://ixa2.si.ehu.eus/convai/maria-bot/).
- Modelo de QA monolingüe para inglés de dominio científico. Este modelo está basado en el modelo de lenguaje SciBERT, que es un modelo BERT de inglés pero entrenado en un corpus grande de textos científicos, y hemos utilizado los conjuntos de preguntas y respuestas SQuAD y QuAC de dominio general. Hemos utilizado este modelo para la implementación de un sistema de QA que busca respuestas a las preguntas en inglés de los expertos relacionados con la enfermedad COVID-19 y el virus SARS-CoV-2 analizando cientos de miles de artículos científicos. Este sistema ha sido premiado en el "COVID-19 Open Research Dataset Challenge (CORD-19)" que organiza el "Office of Science and Technology Policy" de la Casa Blanca de Estados Unidos (https://tinyurl.com/y6ulvffa).
- Modelo de QA multilingüe (inglés, español, euskera) que es capaz de responder preguntas sobre historiales clínicos de los pacientes. Este modelo está basado en el modelo de lenguaje IXAmBERT. Para ajustar este modelo hemos utilizado diferentes tipos de conjuntos de datos. Por un lado, hemos utilizado QuAC, SQuAD y SQuAD-es (la versión traducida al español del anterior), que son conjuntos de preguntas y respuestas de dominio general. Y por otro lado, para ajustar el modelo de QA al dominio de historias clínicas, hemos utilizado la colección emrQA, que contiene 400.000 pares de pregunta-respuesta en inglés sobre historias clínicas.
- Modelo de QA multilingüe que es capaz de responder a preguntas sobre manuales técnicos en español de una empresa privada. Este modelo está basado en el modelo de lenguaje XLM-RoBERTa. Para ajustar este modelo, primero hemos utilizado los conjuntos de datos de dominio

general QuAC, SQuAD y SQuAD-es. Después hemos ajustado más el modelo a la tarea utilizando un conjunto de datos de tamaño reducido, pero del dominio y en español proporcionado por la empresa interesada.

Relacionado también con los sistemas de QA, también hemos creado un modelo multilingüe (euskera, español e inglés) que mide la similitud semántica entre 2 preguntas. Este modelo se basa en el modelo Sentence-BERT (SBERT) que es un modelo capaz de crear representaciones vectoriales de frases en inglés, que posteriormente pueden compararse entre ellas, por ejemplo, utilizando similitud coseno. Basándonos en este modelo, hemos creado un modelo multilingüe que sirve para crear representaciones de frases en diferentes idiomas, entre otros, euskera y español. Utilizando este modelo multilingüe como motor, y aprovechando toda la información que suele haber en una página web de FAQs, hemos creado un sistema de QA que responde a la pregunta planteada por un usuario, devolviendo las preguntas (y respuestas) más similares a esa pregunta que se encuentran en el FAQ.

3.7. Comparación empírica del rendimiento de los modelos del lenguaje

Se ha comparado empíricamente el rendimiento de modelos de lenguaje multilingües como multilingual BERT or XLM-RoBERTa con respecto a modelos de lenguaje monolingües para diversos idiomas y tareas. Por ejemplo, para tareas de etiquetado secuencial como el Reconocimiento de Entidades Nombradas, lematización, POS tagging, análisis de opiniones y detección de metáfora. En general, se ha comprobado que los modelos de lenguaje monolingües, incluyendo aquellos entrenados en el marco de DeepText, obtienen mejores resultados para todas las tareas. Esta tendencia se acentúa en tareas de evaluación con textos de redes sociales como puede ser Twitter. Así, una exhaustiva evaluación en euskera y castellano en el marco de la tarea de evaluación VaxxStance@IberLEF 2021 (https://vaxxstance.github.io/) organizada por investigadores de DeepText, demuestra la clara superioridad de los modelos monolingües desarrollados en DeepText.

3.8. Natural Language Inference (NLI)

En primer lugar, se ha propuesto utilizar modelos de lenguaje preentrenados en la tarea de NLI para resolver la extracción de relaciones entre entidades. La aproximación propuesta se basa en en uso de templetas y técnicas de promp-learning para generar hipótesis sobre relaciones y hacer la clasificación según los modelos de lenguaje rentrenados en la tarea de NI . Este trabajo fue publicado en la conferencia de gran prestigio EMNLP. Siguiendo esta linea de trabajo, se ha re-adaptado la aproximación para acometer las tareas de extracción de eventos (usando múltiples fuentes para entrenar los modelos de lenguaje) y otras tareas de extracción de información. El uso de modelos de lenguaje se demostrado que útil para obtener resultados que mejoran el estado del arte varias tareas de clasificación e inferencia. En segundo lugar, se han investigado estrategias efectivas para la

anotación de ejemplos corrección respuestas de estudiantes mediante el uso de modelos de lenguaje pre-entrenado y reentrenados en la tarea de NLI. Finalmente, se ha investigado el uso de modelos lenguajes multilingües en la tarea de la extracción de relaciones temporales en escenarios de pocos datos anotados y distribuciones de clases muy desbalanceadas. Se ha demostrado que la combinación de técnicas de aumento de datos y modelos grandes de lenguaje es capaz de mitigar los efectos del desbalanceo y la escasez de datos.

3.9. Experimentos en el dominio médico

Informes clínicos y autopsias verbales

Hemos trabajado con informes clínicos (texto narrativo no estructurado) en castellano con objeto de codificarlos según la Clasificación Internacional de Enfermedades (CIE-10) en una aproximación de clasificación multi-etiqueta con BERT con distintas estrategias: 1) Emplear un modelo BERT implementando una cabeza de clasificación multi-etiqueta e incorporando conocimiento sobre la jerarquía del CIE en la cabeza de clasificación mediante una arquitectura multi-task, donde cada task se corresponde a un nivel de la jerarquía. 2) Implementar un módulo de atención multi-etiqueta por cada código CIE sobre un modelo BERT Multi-lingual aplicado a Español, Sueco e Inglés. Por otra parte, hemos trabajado con autopsias verbales (narrativa libre no estructurada en entrevistas a familiares) corroborando la utilidad de los modelos dotados de mecanismos de atención como sistema de ayuda a la clasificación de la causa de muerte por parte de expertos.

Del mismo modo, se ha estudiado como se puede resumir un informe médico mediante las entidades médicas existentes en el mismo. Primero se extraen las entidades médicas (trastornos, tratamientos y tests) de los informes mediante el uso de Bio+ClinicalBeRT con fine-tunning. Los informes se representan mediante embeddings de frases pre-entrenados. Por ejemplo se usa allenai-specter para el dominio científico y LaBSE (Language-agnostic BERT Sentence Embedding) como modelo multilingüe. Se ha participado en TREC 2021 donde se pretende buscar los ensayos clínicos más relevantes para un tópico concreto, usando la similitud de coseno para buscar esos informes médicos más similares a uno concreto (tópico).

Entidades médicas

Se ha estudiado el aprendizaje federado de entidades médicas. En este caso, se aprenden entidades médicas en inglés con distintos corpus de la tarea i2b2, primero de manera centralizada, y luego mediante federación. Para el reconocimiento de entidades se experimenta con BERT y un corpus genérico para el fine-tuning; con Bio_ClinicalBERT+ corpus genérico, y con Bio_ClinicalBERT y 5 conjuntos de textos de i2b2 con diferentes tipos de entidades. Estos experimentos se realizan de manera centralizada. Posteriormente se federan los dos sistemas con mejores resultados y se realizan experimentos cambiando el número de silos, número de comunicaciones y número de

epochs. En un futuro, se investigará como aplicar estas mismas técnicas para entidades médicas en euskera y castellano.

Resultados Percibidos por los Pacientes

Se ha trabajado en la detección de Resultados Percibidos por los Pacientes (PROs - Patient Reported Outcomes) en artículos biomédicos. Este trabajo se ha llevado a cabo en colaboración con BiblioPRO (https://www.bibliopro.org/). Tras una aproximación de clasificación de artículos de PubMed con instrumentos PRO o sin instrumentos PRO, se ha trabajado en el uso del reconocimiento de entidades médicas tipo PRO mediante el uso de modelos pre-entrenados BERT y posterior fine-tuning.

3.10. Aprendizaje y enseñanza del euskera

Hemos adaptado modelos de tipo BERT tanto monolingües como multilingües al dominio del aprendizaje y enseñanza de idiomas. El objetivo es la calificación automática del nivel de euskera (B1/B2/C1/C2) de redacciones de estudiantes que se presentan a exámenes oficiales, y para ello se han entrenado modelos BERT en euskera (BERTeus, roBasquERTa, IXAmBERT, XLM) para la tarea de clasificación de texto. Pese al reducido volumen del corpus de entrenamiento (800 textos cortos), el desequilibrio entre el tamaño de cada clase y presencia de errores ortográficos, los modelos entrenados presentan resultados prometedores.

3.11. Tareas multimodales

Los experimentos se han centrado en utilizar el conocimiento implícito que tienen los modelos de lenguaje para tareas multimodales como en "Outside Knowledge Visual Question Answering", donde el conocimiento científico o el sentido común (entre otros tipos de conocimiento) tienen relevancia para poder resolver la tarea. Para ello, se han re-entrenado tanto Transformers codificadores (BERT y DeBERTa) como Transformers generativos (GPT-2, BART y T5). Hemos llegado a re-entrenar modelos de hasta 11 mil millones de parámetros (T5-11B), viendo la importancia del tamaño o capacidad de la red para codificar este tipo de información. También se ha analizado el uso de "in-context learning" en modelos generativos como GPT-2 de distinto tamaño con el mismo fin, viendo como mayores modelos tienen una capacidad mayor para adaptarse a nuevas tareas. Por último, se ha empezado a explorar la verbalización de grafos de conocimiento para dar dicha información de manera explícita a estos modelos de lenguaje. En un futuro, se estudiará como utilizar este conocimiento implícito en euskera y castellano.

3.12. Mejora de las representaciones

Se ha fine-tuneado un modelo XLMR para la tarea Masked Language Modeling (MLM) con la intención de mejorar las representaciones para el par de idiomas ingles y árabe. Para ello se a recopilado un corpus paralelo y se ha fine-tuneado el modelo recorriendo el corpus linea a linea. Cada linea costa de un texto en ingles y su traducción al árabe, por lo que el modelo debe adivinar las palabras enmascaradas dado el contexto original o su traducción. La finalidad es enriquecer las representaciones correspondientes a este par de idiomas gracias a al corpus paralelo. Se preve estudiar la misma aproximación para pares que contengan el castellano o euskera.

Publicaciones

A continuación, viene el listado de publicaciones de diferentes ámbitos en los que se plasma (y se describe con mayor detalle) la labor llevada a cabo en las diferentes secciones previas. Todas las publicaciones se corresponden con conferencias y workshops y han sido revisadas por pares previa a su aceptación.

- *GAMES: Automatic generation of metadata and multimedia content for media and archives in Basque.* Aitor Álvarez, Ander González-Docasal, Aitor García Pablos, Elena Zotova, Montserrat Cuadros, Haritz Arzelus, Alaitz Artolazabal, Joxe Rojas, Josu Azpillaga, Iban Arantzabal (Project paper in SEPLN2021)
- Ander Barrena, Aitor Soroa, Eneko Agirre. Towards Zero-Shot Cross-Lingual Named Entity Disambiguation. Expert Systems with Applications 2021
- Blanco A., Pérez A., Casillas A, Cobos D. (2021) Extracting Cause of Death From Verbal Autopsy With Deep Learning Interpretable Methods. IEEE Journal of Biomedical Health Informatics 25(4): 1315-1325 (2021)
- Blanco A., Remmer S., Pérez A., Dalianis H., Casillas A. (2021) On the Contribution of Per-ICD Attention Mechanisms to Classify Health Records in Languages with Fewer Resources than English. RANLP 2021: 165-172
- Blanco A. Pérez A., Casillas A. (2020) Extreme Multi-Label ICD Classification: Sensitivity to Hospital Service and Time. IEEE Access 8: 183534-183545 (2020)
- Chung Y.L., Guerini M., Agerri R. (2021). Multilingual Counter Narrative Type Classification. In Argument Mining 2021.
- Joseba Fernandez de Landa & Rodrigo Agerri (2021): Social analysis of young Basque-speaking communities in twitter, Journal of Multilingual and Multicultural Development, DOI: 10.1080/01434632.2021.1962331.
- Garcıa-Pablos, A., Pérez, N., & Cuadros, M. (2021). Vicomtech at eHealth-KD Challenge 2021: Deep Learning Approaches to Model Health-related Text in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*.

- García-Pablos, A., Perez, N., & Cuadros, M. (2021). Vicomtech at MESINESP2: BERT-based Multi-label Classification Models for Biomedical Text Indexing. In BioASQ 2021.
- González-Docasal, A., García-Pablos, A., Arzelus, H., & Álvarez, A. (2021). AutoPunct: A BERT-based Automatic Punctuation and Capitalisation System for Spanish and Basque. *Procesamiento del Lenguaje Natural*, *67*, 59-68.
- Otegi A., San Vicente I., Saralegi X., Peñas A., Lozano B., Agirre E. (2022) Information retrieval and question answering: A case study on COVID-19 scientific literature Knowledge-Based Systems, Volume 240. https://doi.org/10.1016/j.knosys.2021.108072
- Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, Eneko Agirre. Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing 2021.
- Ainhoa Serna, Aitor Soroa, Rodrigo Agerri. Applying Deep Learning Techniques for Sentiment Analysis to Assess Sustainable Transport. Sustainability. 2021; 13(4):2397. https://doi.org/10.3390/su13042397
- Zotova E., Agerri R., Rigau G. (2021). Semi-automatic generation of multilingual datasets for stance detection in Twitter. Expert Systems with Applications, 170 (2021). JCR: 5.452 (Q1). https://doi.org/10.1016/j.eswa.2020.114547
- Zotova, E., Garcıa-Pablos, A.,& Cuadros, M. (2021). Vicomtech at MEDDOPROF: Automatic Information Extraction and Disambiguation in Clinical Text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*.