UNIMER

Tool

Koldo Gojenola Galletebeitia

Arantza Casillas Rubio

Owen Sanchez trigueros

Edgar Andrés santamaría

Documentación UNIMER

Formatos para el reconocimiento de entidades médicas

Nuestro objetivo es unificar las salidas de distintas herramientas MER en un solo formato. Como paso previo a definir un formato nuevo o extender un formato existente, hemos revisado los siguientes formatos. Algunos formatos son genéricos (e.g.: XML, JSON) y otros son específicos del dominio del NLP (e.g.: ConLL-U, Metamap).

Formatos específicos del NLP

A continuación presentamos una serie de formatos diseñados para tareas relacionadas con el lenguaje.

ConLL-U (https://universaldependencies.org/format.html)

ConLL-U emplea un formato que separa los campos en columnas. En la documentación oficial encontramos los siguientes campos:

- ID: Word index, integer starting at 1 for each new sentence; may be a range for multi word tokens; may be a decimal number for empty nodes (decimal numbers can be lower than 1 but must be greater than 0).
- FORM: Word form or punctuation symbol.
- LEMMA: Lemma or stem of word form.
- UPOS: Universal part-of-speech tag.
- XPOS: Language-specific part-of-speech tag; underscore if not available.
- FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
- HEAD: Head of the current word, which is either a value of ID or zero (0).
- DEPREL: Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
- DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.
- MISC: Any other annotation.

Ejemplo (la primera fila no se escribe en el fichero):

```
ID FORM  LEMMA UPOS  XPOS FEATS                                HEAD DEPREL DEPS
MISC
1  They  they  PRON  PRP  Case=Nom|Number=Plur                 2    nsubj
2:nsubj|4:nsubj
2  buy   buy   VERB  VBP  Number=Plur|Person=3|Tense=Pres 0    root   0:root
3  and   and   CONJ  CC   _                                    4    cc     4:cc
4  sell  sell  VERB  VBP  Number=Plur|Person=3|Tense=Pres 2    conj
0:root|2:conj
5  books book  NOUN  NNS  Number=Plur                         2    obj
2:obj|4:obj
6  .     .     PUNCT .    _                                    2    punct  2:punct
```

Pros:

- Permite relaciones (campo DEPS).
- Sencillo de procesar y leer.
- Fácilmente extensible añadiendo las columnas que necesitemos.

Brat standoff (.ann) (https://brat.nlplab.org/standoff.html)

Se trata de un formato más sencillo que captura los siguientes tipos de elementos.

- T: text-bound annotation
- R: relation
- E: event
- A: attribute
- M: modification (alias for attribute, for backward compatibility)
- N: normalization [new in v1.3]
- #: note

Ejemplo:

```
T1 Organization 0 4              Sony
T2 MERGE-ORG 14 27               joint venture
T3 Organization 33 41            Ericsson
E1 MERGE-ORG:T2 Org1:T1 Org2:T3
T4 Country 75 81                 Sweden
R1 Origin Arg1:T3 Arg2:T4
```

Pros:

- Es sencillo, cuenta con tres columnas: tipo de elemento, el hallazgo (por ejemplo, si se trata de una entidad, el tipo de entidad) junto a los offsets y la palabra en sí.
- Existen herramientas de visualización (aunque si extendemos este formato, probablemente dejaría de ser válido como *input* de las mismas).

Cons:

- No es tan completo ni extensible como ConLL-U (al menos esa es nuestra impresión *a priori*).

MetaMap (https://metamap.nlm.nih.gov/)

MetaMap es una herramienta que extrae información médica de los textos. En concreto, *mapea* la información del texto a conceptos médicos, identificando tipos semánticos a partir de distintas fuentes.

Por defecto, usa el formato propio `Fielded\_mmi\_output` (https://metamap.nlm.nih.gov/Docs/README_MetaMapLite_3.6.2rc5.html) que muestra un ranking de todos los mapeos asignados al texto (los ejemplos mostrados a continuación siguen este formato). Dicho esto, acepta todo tipo de formatos de salida, desde JSON a brat.

Los tipos de filas se separan en MMI (MetaMap Indexing), AA (acrónimos y abreviaciones) y UDA (acrónimos y abreviaciones definidos por el usuario.

MMI (MetaMap Indexing) information:

En la documentación oficial (https://metamap.nlm.nih.gov/Docs/MMI_Output_2016.pdf) encontramos los siguientes campos para MMI.

1. ID–Unique identifier used to identify text being processed.
2. MMI–Always MMI.
3. Score–MetaMap Indexing (MMI) score with a maximum score of 1000.00. The higher the score, the greater the relevance of the UMLS concept.
4. UMLS Concept Preferred Name.
5. UMLS Concept Unique Identifier (CUI).
6. Trigger Information–Comma separated sextuple showing what triggered MMI to identify this UMLS concept. (UMLS Concept, loc, locPOS, text, POS and Negation flag). This is used to index UMLS Thesaurus data (https://www.nlm.nih.gov/research/umls/index.html).
7. Location–Summarizes where UMLS concept was found.
8. Positional Information–Semicolon-separated list of positional-information terns, showing StartPos, slash (/), and Length of each trigger identified in the Trigger Information field. (as many semicolon-separated chunks of positional information as there are sextuples of trigger information)
9. Treecode(s)–Semicolon-separated list of any MeSH treecode(s)

Ejemplo:

```
24119710|MMI|170.37|Effect|C1280500|[qlco]|["effects"-ti-1-"Effects"-noun-0]|
TI|21/7|
24119710|MMI|3.44|Various patch test
substance|C0440102|[irda]|["Various"-ab-1-"various"-adj-0]|AB|322/7|
5538822|MMI|2.05|Basal Cell|C0596155|[cell]|["basal cells"-ab-41-"basal
cells"-noun-0]|AB|7059/5,7073/5|
```

AA (Acronym and Abbreviation) information:

En la documentación oficial encontramos los siguientes campos para AA.

1. ID–Unique identifier used to identify text being processed.
2. AA –Always "AA".
3. Short form–The short form of the acronym/abbreviation.
4. Long form–The long form or expansion of the acronym/abbreviation.
5. # of tokens in short form–The number of tokens (including whitespace tokens) in the short form.
6. # of characters in short form–The number of characters in the short form.
7. # of tokens in long form–The number of tokens (including whitespace tokens) in the long form.
8. # of characters in long form–The number of characters in the long form.
9. Positional information of short form–The starting position of the short form followed by a colon (":") followed by the character length of the short form.

Ejemplo:

```
23074487|AA|FY|fiscal years|1|2|3|12|9362:2
23074399|AA|DORs|diagnostic odds ratios|1|4|5|22|8926:4
17342196|AA|PCBs|polychlorinated biphenyls|1|4|3|25|2304:4
```

UDAs (user-defined acronyms and abbreviations)

MetaMap también puede incluir información sobre acrónimos y abreviaciones por el usuario (UDAs).

```
00000000|UA|CHOP|Cyclophosphamide, Hydroxydaunomycin, Oncovin &
Prednisolone|1|4|11|59|8:4
```

Pros: * Configurable output. * Useful additional information for Clinical Area. * Easily integrable output possibilities (.ann and .json)

Cons: * Positional information is not standardized.

There exists a lack of standardization in offset definition, this could be easily fixed using BRAT output format.

CTakes (https://ctakes.apache.org/)

La herramienta de extracción de información médica cTAKES usa una estructura concreta en formato XML como salida, similar a una base de datos relacional, de nombre XMI (XML Metadata Interchange) (https://www.omg.org/spec/XMI/2.5.1/PDF).

La herramienta hace las siguientes tareas:

1. Tokenizado.
2. Normalización.
3. POS.
4. Shallow parsing.
5. MER.
6. Detección de negación.

Pros:

- La salida incluye información lingüística adicional (e.g.: negación).
- La salida puede introducirse en una base de datos relacional directamente.

Cons:

- La información posicional no está estandarizada.
- La salida no es configurable.

En este caso, el formato no es fácilmente integrable con otras estructuras de datos (al menos *a priori*). Además de los beneficios del sistema (alto rendimiento), parece razonable en contraste con las posibilidades actuales de los sistemas contemporáneos.

IOB (Inside Outside Beginning)

IOB es un formato de etiquetado de *tokens* para tareas de *chunking*. En concreto, IOB-2 cuenta con tres tipos de etiquetas:

- I-: etiqueta dentro de un *chunk*.
- O-: etiqueta fuera de cualquier *chunk*.
- B-: etiqueta que da comienzo al *chunk*.

A continuación se muestra un ejemplo.

```
Alex       B-PER
is         O
going      O
to         O
Los        B-LOC
Angeles    I-LOC
in         O
California B-LOC
```

Pros: * Sencillo de interpretar. * Sencillo de procesar.

Cons: * No soporta metadatos como *offsets*, nivel de confianza de la asignación NER, etc. * Tiene que ser extendido para soportarlos. * La etiqueta O no aporta nada de información, podría omitirse.

KAF (https://adimen.si.ehu.es/~rigau/publications/gl09-kaf.pdf)

Se trata de un formato XML especialmente diseñado para establecer la interoperabilidad semántica del conocimiento y del lenguaje que expresa este conocimiento. Sigue una estructura en capas: texto, términos, *chunks* y capas de dependencia. Es compatible con varios lenguajes y cuenta con unas reglas claras para componer el formato.

Pros:

- Al ser una variante de XML (descrita más adelante), cuenta también con sus ventajas.
- Requiere un menor procesamiento para ser aprendido por máquinas.

Cons:

- Requiere que el esquema del diseño encaje con nuestros requerimientos.

El formato resulta ideal para la interacción humano-sistema, pero no deja de presentar dificultades para encajar toda la información que necesitamos en nuestro formato.

Formatos genéricos

Los formatos genéricos no están pensados para recoger información semántica, pero son fácilmente adaptables para recoger todo tipo de información, y, además, todo lenguaje de programación cuenta con librerías para facilitar la lectura y la escritura de los mismos.

XML (https://www.w3.org/standards/xml/core)

XML es un formato simple basado en texto plano para la representación de información estructurada. La información se estructura usando etiquetas y descriptores.

Pros:

- Compatible con sistemas externos.
- compatible con cualquier lenguaje de programación.
- Quick find, update and delete.
- Búsquedas, actualizaciones y borrados rápidos.
- Fácil visualización.
- Permite hacer esquemas adaptables.

Cons:

- Tiene que ser previamente procesado para poder ser *machine learnable*.
- Requiere que el esquema del diseño encaje con nuestros requerimientos.

El formato resulta ideal para la interacción humano-sistema, pero no deja de presentar dificultades a la hora de adaptarlo para esta tarea de NLP y ser empleado directamente como entrada de un sistema ML (de una *pipeline*, por ejemplo).

JSON

Consideramos JSON una alternativa más conveniente que XML porque para nuestra tarea es tan válido como este (XML soporta más tipos de datos, pero sólo usaremos texto y números), siendo más sencilla de leer y escribir.

Por ejemplo, las herramienta NER de SpaCy acepta este tipo de JSON como formato de datos de entrenamiento (https://spacy.io/api/data-formats#json-input):

```
[{
    "id": int,                      # ID of the document within the corpus
    "paragraphs": [{                # list of paragraphs in the corpus
        "raw": string,              # raw text of the paragraph
        "sentences": [{             # list of sentences in the paragraph
            "tokens": [{            # list of tokens in the sentence
                "id": int,          # index of the token in the document
                "dep": string,      # dependency label
                "head": int,        # offset of token head relative to token
index
                "tag": string,      # part-of-speech tag
                "orth": string,     # verbatim text of the token
                "ner": string       # BILUO label, e.g. "O" or "B-ORG"
            }],
            "brackets": [{          # phrase structure (NOT USED by current
models)
                "first": int,       # index of first token
                "last": int,        # index of last token
                "label": string     # phrase label
            }]
        }],
        "cats": [{                  # new in v2.2: categories for text
classifier
            "label": string,        # text category label
            "value": float / bool   # label applies (1.0/true) or not
(0.0/false)
        }]
    }]
}]
```

Este es un ejemplo de SpaCy:

```
[{
    "id": 42,
    "paragraphs": [
      {"raw": "In an Oct. 19 review of \"The Misanthrope\" at Chicago's
Goodman Theatre (\"Revitalized Classics Take the Stage in Windy City,\"
Leisure & Arts), the role of Celimene, played by Kim Cattrall, was mistakenly
attributed to Christina Haag. Ms. Haag plays Elianti.",
        "sentences": [
          {
            "tokens": [
              {
                "head": 44,
                "dep": "prep",
                "tag": "IN",
                "orth": "In",
                "ner": "O",
                "id": 0
              },
              {
                "head": 3,
                "dep": "det",
                "tag": "DT",
                "orth": "an",
                "ner": "O",
                "id": 1
              }
              ...
            ],
            "brackets": [
              {
                "first": 2,
                "last": 3,
                "label": "NML"
              },
              {
                "first": 1,
                "last": 4,
                "label": "NP"
              }
            ]
            ...
          }
        ]
      }]}]
```

Pros:

- Fácil de leer.
- Fácil de extender.
- Librerías en todas las herramientas

Cons:

- Supondría crear un formato nuevo (si partimos de ConLL-U o brat standoff, por ejemplo, tendríamos ya una estructura base que después extenderemos al resto de herramientas).

Conclusiones

Tras un estudio de formatos y diversas reuniones se estableció IOB como formato utilizado en los entrenamientos de los sistemas "Deep", por otro lado se estableció que la entrada al sistema debía soportar un fichero compuesto de archivos ".txt" y proveer un fichero de salida compuesto de los correspondientes archivos ".conllv2" para cada texto de entrada, este formato se explica en profundidad en el primer anexo de los apéndices. Este formato derivado de ".conll" permite la integración de información generada por FreelingMed, DeepMER y además DeepREL Estos sistemas se exponen teóricamente en la siguiente sección .

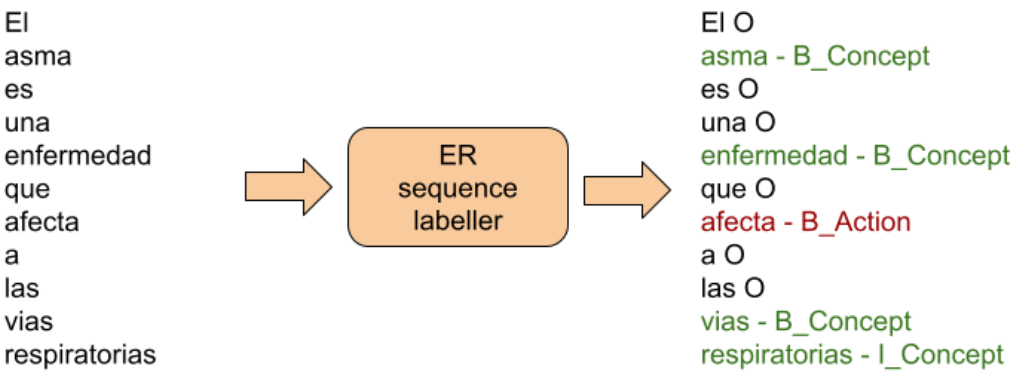
El sistema

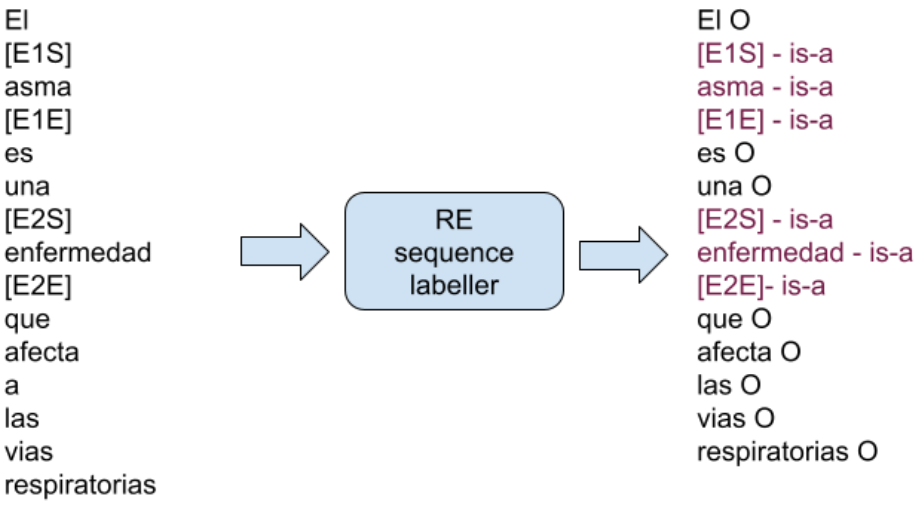
Cómo funciona

Los principios básicos de FLAIR se explican en IXA-NER-RE at eHealth-KD Challenge 2020,. En el artículo se explica el funcionamiento general de un sistema que fundamenta la manera en que se ha desarrollado Uimer, asimismo se tratan detalles y técnicas que ayudan a comprender las decisiones tomadas en el desarrollo del proyecto, y los principios sobre transformers se especifican en IXA at eHealth-KD Challenge 2021 . Los resultados obtenidos a lo largo del desarrollo se encuentran en las tablas provistas al final de cada submódulo Deep, asimismo se añade el documento que explica el formato de los datos utilizados (.conllv2) como primer apéndice al final del documento.

Descripción

La herramienta es un sistema modular en dos pasos, primero se extraen las entidades médicas y las negaciones, para ello se procesan las historias clínicas a formato IOB y se aprende automáticamente a predecir las secuencias de eventos. Finalmente se genera un escenario de entrenamiento para la extracción de relaciones, en este caso se crean todos los posibles pares de entidades y se aprende a inferir aquellas que tienen relación.

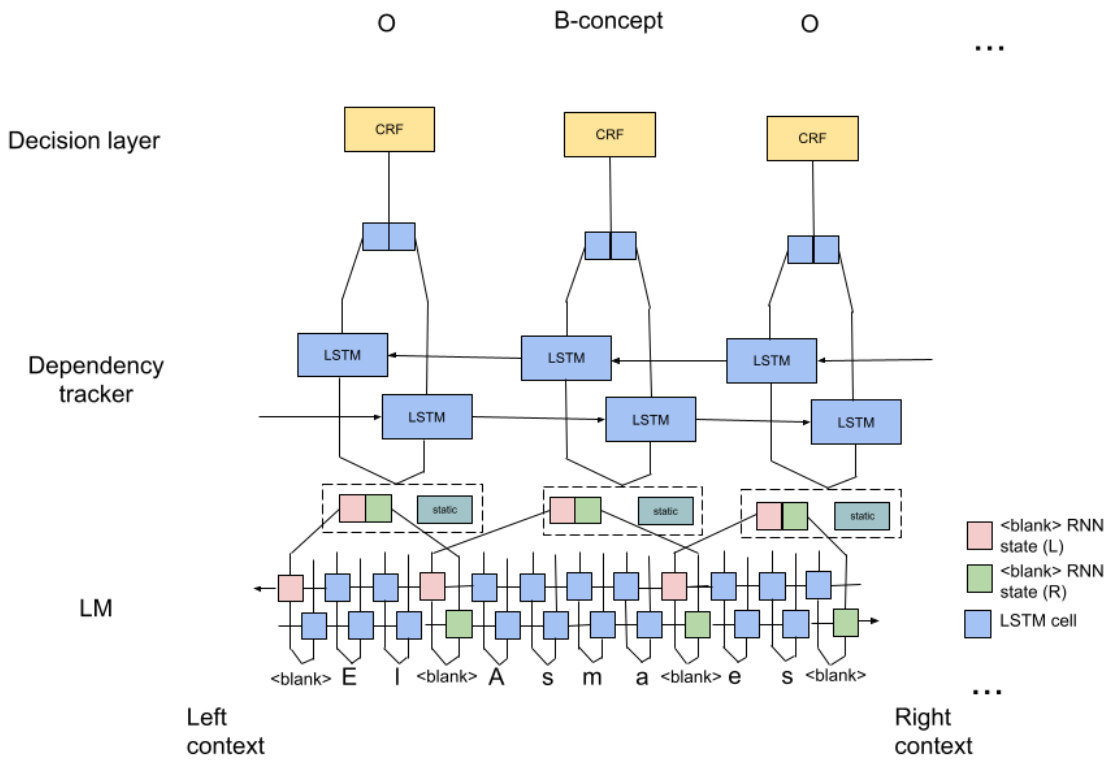




FLAIR

El sistema FLAIR (https://github.com/flairNLP/flair) es capaz de aprender representar las palabras y su contexto, esta representación es muy útil para dominios específicos como el clínico. Finalmente se aprenden las secuencias de eventos gracias a la capa CRF. En este caso y según los experimentos llevados a cabo es la arquitectura seleccionada para la detección de Entidades Médicas y Negaciones en textos clínicos

Cabe destacar la capacidad de combinar diferentes tipos de representaciones para palabras, en la siguiente figura se puede apreciar la composición, en la capa LM vemos cómo se calcularán las representaciones "character based", posteriormente se concatenan con los "embeddings estáticos", y finalmente en la capa Dependency Tracker se calculan las representaciones contextuales "flair embeddings". Estas combinaciones son configurables pudiendo omitirlas o incluso sumar varios componentes en la misma capa obteniendo una media de ellos .



Transformers

El sistema Transformers (https://huggingface.co/transformers/) es capaz de aprender representar las palabras y su contexto atendiendo a las palabras más relevantes, esta representación es muy útil para dominios específicos como el clínico con la contraparte de requerir gran cantidad de ejemplos para un correcto funcionamiento. Finalmente se aprenden las secuencias de eventos gracias a "fully connected layers". En este caso y según los experimentos llevados a cabo es la arquitectura seleccionada para la detección de relaciones entre conceptos clínicos.

Reconocedores de entidades médicas (MER)

FreeLingMed

Descripción de la herramienta

FreeLingMed es una librería de análisis de lenguaje clínico que ofrece funcionalidades relacionadas con: morphological analysis, named entity detection, PoS-tagging, parsing,

Word Sense Disambiguation by Semantic Role Labelling, para una amplia variedad de lenguajes como Inglés y Castellano entre otros.

Formato de entrada

```
1  cáncer                 "El paciente no presenta cáncer ni anemia"
2  anemia                 "El paciente no presenta cáncer ni anemia"
3  anemia                 "El paciente no presenta cáncer pero si tiene
anemia"
4  infarto                "Oclusion y estenosis otra arteria espec no infarto
cerebral"
5  infarto cardiovascular "infarto cardiovascular por oclusion y estenosis
otra arteria espec no infarto cerebral"
6  infarto cerebral       "infarto cardiovascular por oclusion y estenosis
otra arteria espec no infarto cerebral"
7  dolor cronico          "paciente sin dolor cronico"
```

Formato de salida

```
1  cancer                 "El paciente no presenta cáncer ni anemia"
Negated  negPhrases
2  anemia                 "El paciente no presenta cáncer ni anemia"
Negated  negPhrases
3  anemia                 "El paciente no presenta cáncer pero si tiene
anemia"                                  Affirmed NONE

…
7  dolor cronico          "paciente sin dolor cronico"
Negated  negPhrases
```

También muestra por STDOUT:

```
3 - 6
3 - 6
3 - 6
…

2 - 3
```

DeepMER

Descripción de la herramienta

En este caso se provee un sistema basado en redes neuronales recurrentes (RNNs) y desarrollado mediante el framework Flair. Esta tecnología nos permite aplicar un completo Modelo de Lenguaje (LM) mediante la combinación de vectores: de palabra, de contexto y de carácter. En la capa superior del LM se encuentra una arquitectura BiLSTM + CRF para realizar la clasificación por palabra (token Classification). De esta manera el sistema establece los parámetros para mapear una secuencia de palabras a una secuencia de etiquetas

Formato de entrada

Existen dos escenarios principales, el primero consiste en "aprender" cómo dada la secuencia de palabras predecimos secuencia de etiquetas de etiquetas (Train schema), en el segundo escenario pretendemos dada una secuencia de palabras predecir la secuencia de etiquetas (Prediction scheme).

Train schema

```
El -
paciente -
no -
presenta -
cancer B-Grp_Enfermedad
ni -
anemia B-Grp_Enfermedad
```

Prediction schema

```
El -
paciente -
no -
presenta -
cancer -
ni -
anemia -
```

Formato de salida

Existen dos escenarios principales en la herramienta, el primero consiste en "aprender" cómo dada la secuencia de palabras predecimos secuencia de etiquetas de etiquetas utilizando los datos provistos (Train schema), en este caso se consigue el artefacto "modelo" capaz de desempeñar la tarea afinado, en el segundo escenario pretendemos dada una secuencia de palabras (Prediction scheme) predecir la secuencia de etiquetas, en este caso el resultado es un fichero .tsv en el que la framework Flair vuelca los resultados sobre tres columnas (words, gold, pred).

Prediction output

```
El - -
paciente - -
no - -
presenta - -
cancer B-Grp_Enfermedad B-Grp_Enfermedad
ni - -
anemia B-Grp_Enfermedad B-Grp_Enfermedad
```

Resultados experimentales MER

En esta sección se exponen los resultados experimentales realizados para determinar la combinación paramétrica y el sistema más adecuados para la tarea Medical Entity Recognition, se exponen 7 experimentos (exp 1- 7) donde el primero fué realizado utilizando la arquitectura transformers, y los siguientes mediante la arquitectura FLAIR.

Para los entrenamientos se utilizó el corpus Berdeak (Train, Dev y Test), este sigue el formato (.conllv2), en está tarea se crearon ficheros IOB utilizando las columnas: Hitz-forma (2) y AnnEnt (14), para el caso exp7 se aplicó una función "lowercase" a la columna Hitz-forma (2).

Configuración Flair Embeddings

exp2	WordEmbeddings("es"), FlairEmbeddings("es forward" chars_per_chunk=128), FlairEmbeddings("es_backward" chars_per_chunk=128),
exp3	WordEmbeddings("medical"), FlairEmbeddings("medical_forward" chars_per_chunk=128), FlairEmbeddings("medical_backward" chars_per_chunk=128),
exp4	WordEmbeddings("es"), CharacterEmbeddings(), FlairEmbeddings("es_forward"), FlairEmbeddings("es_backward"),
exp5	WordEmbeddings("medical"), CharacterEmbeddings(), FlairEmbeddings("medical_forward"), FlairEmbeddings("medical_backward"),
exp6 and exp7	WordEmbeddings('es'), WordEmbeddings("wikipedia"), WordEmbeddings('medical'), CharacterEmbeddings(), FlairEmbeddings('medical_forward'), FlairEmbeddings('medical_backward'), FlairEmbeddings('es-forward'), FlairEmbeddings('es-backward'),

Descriptores Flair Embeddings

"es"	Spanish FastText embeddings over News
"es_forward"	Spanish forward contextual embeddings over News
"es_backward"	Spanish backward contextual embeddings over News
"medical"	'/gscratch2/users/igoenaga006/JCRArtikuluak/Datasetak/Baliabideak/Medikuntzako_Bektoreak.gensim'
"medical_forward"	"'/gscratch2/users/igoenaga006/JCRArtikuluak/Datasetak/PharmacoNER2019/resources/taggers/language_model-forward-Finetunning/best-lm.pt

Configuración de las Frameworks

FLAIR	Framework	Transformers	Framework
hidden_size	256	load_best_model_at_end:	true
use_crf	True	metric_for_best_model	f1
learning_rate	0.1	evaluation_strategy	steps
mini_batch_size	30	num_train_epochs	5
train_with_dev	True	per_device_train_batch_size	4
embeddings_storage_mode	'gpu'	per_device_eval_batch_size	4
max_epochs	5	eval_steps	2000
		seed	1

Resultados MER

ID	F1	Precision	Recall	Accuracy
exp1	86,510	85,04	88,03	96,94
exp2	84,24	84,18	84,3	95,61
exp3	87,28	88,29	86,29	97,52
exp4	84,11	83,18	85,07	95,68
exp5	87,29	87,88	86,72	96,49
exp6	88,45	89,35	87,57	96,69
exp7	80,7	79,1	82,36	96,59

Mejores Resultados MER Desglose

exp5	Precision	recall	f1
Calificador	82,22	71,49	76,48
Estructura_Corporal	84,66	88,31	86,45
Grp_Enfermedad	90,16	90,25	90,21
Grp_Medicamento	93,42	90,74	92,06
Procedimiento	0	0	0
Alergia	96,43	100	98,18
exp6	Precision	recall	f1
Calificador	80,85	75,04	77,84
Estructura_Corporal	87,03	88,74	87,87
Grp_Enfermedad	92,82	90,31	91,54
Grp_Medicamento	92,72	91,94	92,33
Procedimiento	0	0	0
Alergia	90,96	89,47	90,21

Comparativa Resultados MER Desglose

IAKES (Lample)	Sarea Berdeak		
	Precision	Recall	F1
Orokorra	87.57	86.09	86.83
Calificador	77.04	72.61	74.76
Enfermedad	90.94	88.82	89.87
Medicamento	93.2	90.23	91.69
Estructura_C	86.75	88.74	87.74
Procedimiento	61.53	18.6	28.57

Reconocedores de Negación

DeepNEG

Descripción de la herramienta

En este caso se provee un sistema basado en redes neuronales recurrentes (RNNs) y desarrollado mediante el framework Flair. Tal y como se ha introducido previamente para DeepMER.

Formato de entrada

En este caso el formato de entrada IOB coincide con lo expuesto en DeepMER, en este caso se utilizan las correspondientes etiquetas a la tarea de negación/ especulación

Train schema

```
El O
paciente O
no O
presenta O
cancer Negacion
ni O
anemia Negacion
```

Prediction schema

```
El O
paciente O
no O
presenta O
cancer O
ni O
anemia O
```

Formato de salida

Tal y como sucede con DeepMER, existe un fichero .tsv en el que la framework Flair vuelca los resultados sobre tres columnas (words, gold, pred)

Prediction output

```
El O O
paciente O O
no O O
presenta O O
cáncer Negacion  Negacion
ni O O
anemia Negacion  Negacion
```

Resultados experimentales NEG

En esta sección se exponen los resultados experimentales realizados para determinar la combinación paramétrica y el sistema más adecuados para la tarea Negation Recognition, se exponen 7 experimentos (exp 1- 7) donde el primero fué realizado utilizando la arquitectura transformers, y los siguientes mediante la arquitectura FLAIR.

Para los entrenamientos se utilizó el corpus Berdeak (Train, Dev y Test), este sigue el formato (.conllv2), en está tarea se crearon ficheros IOB utilizando las columnas: Hitz-forma (2) y Ez/Espek(18), para el caso exp7 se aplicó una función "lowercase" a la columna Hitz-forma (2).

En este caso las configuraciones de FLAIR embeddings y paramétricas de las frameworks coinciden con las expuestas en la sección Resultados experimentales MER, la diferencia reside en los datos utilizados, en este caso la tarea Ez/Espek.

Resultados NEG

ID	F1	Precision	Recall	Accuracy
exp1	89,490	88,09	90,94	99,71
exp2	90,02	94,5	85,94	99,68
exp3	91,75	95,8	88,02	99,74
exp4	87,06	97,18	78,85	99,61
exp5	91,03	91,66	90,42	99,71
exp6	92,02	93,35	90,73	99,74
exp7	91,76	94,29	89,38	99,74

Mejores Resultados NEG Desglose

Exp7	Precision	recall	f1
Especulación	94,44	49,28	64,76
Negación	95,86	91,43	93,59
Negación|Especulación	0	0	0
Exp6	Precision	recall	f1
Especulación	78,79	75,36	77,04
Negación	94,46	92,33	93,39
Negación|Especulación	0	0	0

Comparativa Resultados NEG

Entity	Phrase	Real	NegEx Result	NegEx Span	Flair Result
dolor	intento exéresis pero no lo consigo por dolor	Affirmed	Negated	4 - 7	Affirmed
micro infarto	micro infarto no q eac monovaso	Affirmed	Affirmed	2 - 4	Affirmed
sangre	sangre no fresca en todo el colon mayor cantidad en colon izquierdo con algún coágulo	Affirmed	Affirmed	2 - 14	Affirmed
hemoptisis	hemoptisis leve autolimitada no filiada	Affirmed	Affirmed	4 - 4	Affirmed
broncoespasmo	colonoscopia no se pudo proseguir más allá de 30 cm por broncoespasmo y cianosis	Affirmed	Negated	2 - 13	Affirmed
irritación	no signos de irritación peritoneal	Affirmed	Negated	1 - 4	Affirmed
herida	herida no complicada	Affirmed	Affirmed	2 - 2	Affirmed
hemorroides	tr en urgencias hemorroides no complicadas	Affirmed	Affirmed	5 - 5	Affirmed
dolor	ese dolor hoy no ha cambiado	Affirmed	Affirmed	4 - 5	Affirmed
dolor	dolor a la palpación ligamento lateral externo fascículo anterior	Affirmed	Affirmed	-	Affirmed
dolor	nolotil 18 h si dolor	Affirmed	Affirmed	-	Affirmed
elevación	analítica sin alteraciones significativas salvo elevación dd 24	Affirmed	Negated	2 - 7	Affirmed
epoc	epoc sin tratamiento	Affirmed	Affirmed	2 - 2	Affirmed
hta	hta sin tto farmacológico	Affirmed	Affirmed	2 - 3	Affirmed
urgencia	cuenta urgencia defecatoria sin descanso nocturno	Affirmed	Affirmed	4 - 5	Affirmed
hepatopátia	datos de hepatopatía crónica	Affirmed	Affirmed	-	Affirmed
cicatriz	tórax cicatriz de toracotomía media y dai implantado en zona infraclavicular izda	Affirmed	Affirmed	-	Affirmed
soplo	ac rítmico soplo sistólico en ápex	Affirmed	Affirmed	-	Affirmed
dm	no antecedentes de dm hepatopatía nefropatía ni dislipemia	Negated	Negated	1 - 7	Negated
hepatopatía	no antecedentes de dm hepatopatía nefropatía ni dislipemia	Negated	Negated	1 - 7	Negated
nefropatía	no antecedentes de dm hepatopatía nefropatía ni dislipemia	Negated	Negated	1 - 7	Negated
dislipemia	no antecedentes de dm hepatopatía nefropatía ni dislipemia	Negated	Negated	1 - 7	Negated
fiebre	niega fiebre cefalea dolor torácico o palpitaciones	Negated	Negated	1 - 6	Negated
cefalea	niega fiebre cefalea dolor torácico o palpitaciones	Negated	Negated	1 - 6	Negated

dolor	niega fiebre cefalea dolor torácico o palpitaciones	Negated	Negated	1 - 6	Negated
palpitaciones	niega fiebre cefalea dolor torácico o palpitaciones	Negated	Negated	1 - 6	Negated
endocarditis	reingresa 1 mes después por sdre febril atribuido a itu con bacteriemia por e faecalis descartando se mediante repetidos etts endocarditis	Negated	Affirmed	-	Affirmed
estenosis	eco doppler de tsa hay fibro placas sin estenosis significativas	Negated	Negated	7 - 8	Negated
hta	no hta dm ni otras metabolopatías	Negated	Negated	1 - 5	Negated
dm	no hta dm ni otras metabolopatías	Negated	Negated	1 - 5	Negated
metabolopatías	no hta dm ni otras metabolopatías	Negated	Negated	1 - 5	Negated
soplos	ac rítmico no oigo soplos	Negated	Negated	3 - 4	Negated
lesión	rx columna cervical 2p no evidencia de lesión ósea aguda	Negated	Negated	5 - 9	Negated
soplos	ac rítmico a 90x sin soplos	Negated	Negated	5 - 5	Negated
dolor	al alta afebril excelente estado general no refiere dolor abdominal abdomen blando y depresible salto negativo	Negated	Negated	7 - 15	Negated
lesión	rx caderas ap sin evidencia de lesión ósea aguda	Negated	Negated	5 - 8	Negated
origen	en los ingresos previos se ha descartado un origen tumoral del sangrado tratándose por tanto lo más probable de una hdb secundaria a una lesión vascular benigna angiodisplasias vs divertículos	Negated	Affirmed	-	Affirmed
visceromegalia	abdomen blando y depresible no visceromegalia sin signos de irritación peritoneal	Negated	Negated	5 - 10	Negated
irritación	abdomen blando y depresible no visceromegalia sin signos de irritación peritoneal	Negated	Negated	5 - 10	Negated
hábitos	niega hábitos tóxicos	Negated	Negated	1 - 2	Negated
hábitos	no hábitos tóxicos	Negated	Negated	1 - 2	Negated
iy	cyc no iy	Negated	Negated	2 - 2	Negated
alergias	niega alergias medicamentosas conocidas	Negated	Negated	1 - 3	Negated
romberg	romberg negativo	Negated	Affirmed	-	Affirmed
ulceraciones	orofaringe normal sin ulceraciones	Negated	Negated	3 - 3	Negated
ideas	no ideas de muerte estructuradas	Negated	Negated	1 - 4	Negated
alergias	alergias no conocidas	Negated	Affirmed	2 - 2	Negated
condensaciones	no observó condensaciones ni pinzamientos	Negated	Negated	1 - 4	Negated
pinzamientos	no observó condensaciones ni pinzamientos	Negated	Negated	1 - 4	Negated
Ejemplos inventados o modificaciones sobre el corpus					
tos	no tos pero sí fiebre	Negated	Negated	1 - 3	Negated
fiebre	no tos pero sí fiebre	Affirmed	Affirmed	1 - 3	Affirmed
covid	sí covid no fiebre	Affirmed	Affirmed	3 - 3	Affirmed

fiebre	sí covid no fiebre	Negated	Negated	3 - 3	Affirmed
covid	no covid sí fiebre	Negated	Negated	1 - 3	Negated
fiebre	no covid sí fiebre	Affirmed	Negated	1 - 3	Affirmed
inestabilidad	hdb sin inestabilidad hd en probable relación a diverticulosis colónica en paciente anticoagulado	Negated	Negated	2 - 12	Negated
diverticulosis	hdb sin inestabilidad hd en probable relación a diverticulosis colónica en paciente anticoagulado	?	Negated	2 - 12	Affirmed
calor	en zona glúteo no presenta induración ni calor ni rubor si dolor a la palpación a nivel del glúteo y área paralumbar	Negated	Negated	4 - 21	Negated
rubor	en zona glúteo no presenta induración ni calor ni rubor si dolor a la palpación a nivel del glúteo y área paralumbar	Negated	Negated	4 - 21	Negated
dolor	en zona glúteo no presenta induración ni calor ni rubor sí dolor a la palpación a nivel del glúteo y área paralumbar	Affirmed	Negated	4 - 21	Affirmed
metástasis	no se encuentra ninguna sugerencia de metástasis	Negated	Negated	1 - 6	Negated
metástasis	se descarta metástasis	Negated	Affirmed	0 - 0	Affirmed
origen	en los ingresos previos se ha descartado un origen tumoral del sangrado tratándose por tanto lo más probable de una hdb secundaria a una lesión vascular benigna angiodisplasias vs divertículos	Negated	Affirmed	-	Affirmed
un origen tumoral	en los ingresos previos se ha descartado un origen tumoral del sangrado tratándose por tanto lo más probable de una hdb secundaria a una lesión vascular benigna angiodisplasias vs divertículos	Negated	Affirmed	-	Affirmed
tumor	no tumor pero sí hernia	Negated	Negated		Negated
hernia	no tumor pero sí hernia	Affirmed	Affirmed		Affirmed
hernia	no tumor sí hernia	Affirmed	Negated		Affirmed

Reconocedores de Relaciones

DeepREL

Descripción de la herramienta

En este caso se propone el sistema basado en LMs pre-entrenados y desarrollado mediante el framework Transformers. Esta tecnología nos permite aplicar un completo Modelo de Lenguaje (LM) calculado a través de arquitecturas codificador-decodificador. En la capa superior del LM se establecen capas de codificador-decodificadores para retener la información semántica necesaria para la predicción de secuencias. De esta manera el sistema establece los parámetros para mapear una secuencia de palabras a una secuencia de etiquetas

Formato de entrada

Particularmente en este escenario se requiere la salida del reconocedor de entidades DeepMER, las entidades reconocidas se combinarán cubriendo todas las posibilidades, cada una de estas por separado se agrega com un ejemplo y se diferencia mediante la añadidura de “entity markers”, concretamente: E1S , E1E , E2S y E2E.

Existen dos escenarios principales en la herramienta, el primero consiste en “aprender” cómo dada la secuencia de palabras predecimos secuencia de etiquetas de etiquetas (Train schema), en el segundo escenario pretendemos dada una secuencia de palabras predecir la secuencia de etiquetas (Prediction scheme).

Train schema

```
El -
paciente -
no -
presenta -
[E1S] LOC
cancer LOC
[E1E] LOC
de -
[E2S] LOC
pulmón LOC
[E2E] LOC
```

Prediction schema

```
El -
paciente -
no -
presenta -
[E1S]-
cancer -
[E1E] -
de -
[E2S] -
pulmón -
[E2E] -
```

Formato de salida

Existen dos escenarios principales en la herramienta, el primero consiste en “aprender” cómo dada la secuencia de palabras predecimos secuencia de etiquetas de etiquetas utilizando los datos provistos (Train schema), en este caso se consigue el artefacto “modelo” capaz de desempeñar la tarea afinado, en el segundo escenario pretendemos dada una secuencia de palabras (Prediction scheme) predecir la secuencia de etiquetas, en este caso solo contamos con un único formato de salida similar al propuesto en “.conllv2”:

1	El	O	-	-	-	
2	paciente	O	-	-	-	
3	no	O	-	-	-	
4	presenta	O	-	-	-	
5	cancer	B-Grp_Enfermedad	7	1	Modif	
6	de	O	-	-	-	
7	pulmón	B-Estructura_Corporal	-	-	-	

Resultados experimentales REL

En esta sección se exponen los resultados experimentales realizados para determinar la combinación paramétrica y el sistema más adecuados para la tarea Relation Extraction, se exponen 2 experimentos (exp 1 y 2) donde el primero fué realizado utilizando la arquitectura transformers y , y el segundo mediante la arquitectura FLAIR.

Para el primer experimento (exp1) el corpus Berdeak (Train, Dev y Test), este sigue el formato (.conllv2), se crean ficheros IOB utilizando las columnas: Hitz-forma (2), AnnEnt(14) y AnnErl(17), estas columnas sirven para generar todos los pares de entidades posibles y con los datos generados se evalúa el sistema.

Para el segundo experimento se utilizó un método de preproceso diferente, en este caso la cadena IOB era creada siguiendo un patrón : [direction][number][tag], donde "direction" (| M); "number" (BAT | BI | HIRU | LAU | BOST ….); "tag" (LOC | MODIF | CP). De está manera se etiqueta la entidad a relacionar con la entidad relacionada que se encuentra en "direction", la "number" entidad con el tipo de relación "tag". Por ejemplo podemos ver como cáncer se relaciona con la siguiente entidad a su derecha mediante la etiqueta LOC:

Train schema	Prediction schema
El -	El -
paciente -	paciente -
no -	no -
presenta -	presenta -
cancer BATLOC	cancer -
de -	de -
pulmón -	pulmón -

Resultados REL

ID	F1	Precision	Recall	Accuracy
exp1	86,68	88,17	85,24	99,2
exp2	74,8	73,92	75,7	98,18

Resultados REL Desglose

Exp1	Precision	recall	f1
Causada_por	0,6663	0,5216	0,5852
LOC	0,881	8452	0,8627
Modif	0,9084	0,9127	0,9105
Exp2	Precision	recall	f1
BATCP	0,3889	0,2979	0,3373
BATLOC	0,4795	0,4432	0,4607
BATMODIF	0,7564	0,7195	0,7375
BICP	0	0	0
BILOC	0,4167	0,5102	0,4587
BIMODIF	0	0	0
BOSTLOC	0	0	0
Causada_por_Aurreko_Esaldi_Batean	0	0	0
Causada_por_Hurrengo_Esaldi_Batean	0	0	0
HIRUCP	0	0	0
HIRULOC	0	0	0
LAUCP	0	0	0
LAULOC	0	0	0
LOC_Hurrengo_Esaldi_Batean	0,5161	0,32	0,3951
MBATCP	0,6995	0,772	0,734
MBATLOC	0,8041	0,8372	0,8203
MBATMODIF	0,7582	0,8269	0,7911
MBICP	0	0	0
MBILOC	0	0	0
MBIMODIF	0	0	0
MHIRUCP	0	0	0
MLAUMODIF	0	0	0
Modif_Aurreko_Esaldi_Batean	0	0	0
SEILOC	0	0	0

Conclusiones

En este apartado hemos introducido los diferentes submódulos de la aplicación, estos están preparados para su posible integración con aplicaciones externas:

- /tartalo03/DOTT-HEALTH-PIC/software/freelingmed/freelingmed
- /tartalo03/DOTT-HEALTH-PIC/software/flairmer/flairmer
- /tartalo03/DOTT-HEALTH-PIC/software/unimer/unimer

En el primer modulo contamos con la interfaz para el etiquetado de textos utilizando el sistema FreelingMed, en el segundo para el etiquetado de textos utilizando los sistemas DeepMER y DeepNEG basados en FLAIR, y el último para el etiquetado integrado de FreelingMed, DeepMER y DeepNEG con la opción para añadir los resultados de DeepREL. Las herramientas, su utilización (user API) y la manera de integrarse (developer API) se exponen en la siguiente sección.

Manual de usuario

En está sección se expone el manual de usuario "ready to use" sobre las herramientas alojadas en el servidor "ixa.eus", independientemente de la máquina que utilicemos: mamarro, txorompio, mari, kixmi o traganarru. La única precondición es que contemos con la **GPU 0 a disposición.**

Como paso previo al uso del Software debemos apuntar a la carpeta contenedora de las herramientas:

cd /tartalo03/DOTT-HEALTH-PIC/software

Flairmer

Flairmer debe ejecutarse como un módulo al cual le pasamos dos argumentos "input_dir", con la siguiente estructura:

input_dir/

|_____ doc1.txt

|_____ doc1.txt

|_____ doc1.txt

...

Y el segundo apuntado al directorio de salida, el módulo se ejecutará de la siguiente manera. Primero nos desplazamos a la carpeta flairmer:

cd flairmer/

Después debemos establecer el entorno de la herramienta:

source venv/bin/activate.csh

Y finalmente lanzamos el programa:

python -m flairmer path/to/input_dir/ path/to/output_dir/

El resultado sigue las anotaciones del formato (.conllv2):

output_dir/

|_____ doc1.conll

|_____ doc2.conll

|_____ doc3.conll

…

Los atributos que componen la salida son:

1. **Hitz-zenbakia**: hitzaren posizioa esaldiarekiko.

2. **Hitz-forma**: hitzaren testuko forma.

3. **Offseta**: hitza esaldian ze posiziotan hasten eta bukatzen den adierazten duen tartea.

4. **Dokumentu-izena**: hitza agertzen den dokumentuaren izena.

5. **Esaldi-zenbakia**: hitza dokumentuko zenbatgarren esaldian agertzen den adierazten duen zenbakia.

6. **DeepEnt**: neurona-sare sakonak hitzari esleitu dion entitate motaren etiketa chunkak zehazteko IOB formatuarekin konbinatuta, (B-Grp_Enfermedad, I-Grp_Medicamento, O, B-Alergia).

7. **Ez-Espek**: entitatea eskuzko anotazioan espekulazioa den edo ezeztatuta dagoen adierazten duen etiketa.

Ejemplo:

```
12  las  83-86  Digestivo220161292_19-04-63  SENT108  -  -
13  2  87-88  Digestivo220161292_19-04-63  SENT108  -  -
14  costillas  89-98  Digestivo220161292_19-04-63  SENT108  B-Estructura_Corporal  -
15  y  99-100  Digestivo220161292_19-04-63  SENT108  -  -
16  las  101-104  Digestivo220161292_19-04-63  SENT108  -  -
17  adenopatías  105-116  Digestivo220161292_19-04-63  SENT108  B-Grp_Enfermedad  -
```

Freelingmed

Freelingmed debe ejecutarse como un módulo al cual le pasamos dos argumentos "input_dir", con la siguiente estructura:

```
input_dir/
        |_____ doc1.txt
        |_____ doc1.txt
        |_____ doc1.txt
         ...
```

Y el segundo apuntado al directorio de salida, el módulo se ejecutará de la siguiente manera. Primero nos desplazamos a la carpeta freelingmed:

cd freelingmed/

Después debemos establecer el entorno de la herramienta:

source venv/bin/activate.csh

Adicionalmente debemos establecer las variables de entorno, si tenemos una terminal Bash ($ detrás de nuestro nombre en la línea de comandos):

source set-env-vars.sh

Y si tenemos una terminal Shell(% detrás de nuestro nombre en la línea de comandos):

source /tartalo01/users/eandres011/set-env-vars-freelingmed.sh

Y finalmente lanzamos el programa:

python -m freelingmed path/to/input_dir/ path/to/output_dir/

El resultado sigue las anotaciones del formato (.conllv2):

output_dir/

|_____ doc1.conll

|_____ doc2.conll

|_____ doc3.conll

…

Los atributos que componen la salida son:

1. **Hitz-zenbakia**: Hitzaren posizioa esaldiarekiko.
2. **Hitz-forma**: Hitzaren testuko forma.
3. **Lema**: Hitzaren lema.
4. **FreePOS**: FreelingMedek eman dion kategoria.
5. **Offseta**: Hitza esaldian ze posiziotan hasten eta bukatzen den adierazten duen tartea.
6. **Dokumentu-izena**: Hitza agertzen den dokumentuaren izena.
7. **SnoMot**: SnoMedek eman dion mota.
8. **SnoKod**: SnoMedek eman dion kodea.

Example:

79 para para SPS00 133-137 Digestivo220161292_19-04-63 C0521125 calificador

80 presentación presentación NCFS000 138-150 Digestivo220161292_19-04-63 C0449450 atributo

81 posterior posterior AQ0CS0 151-160 Digestivo220161292_19-04-63 C0205095#C0439781 calificador

82 en en SPS00 161-163 Digestivo220161292_19-04-63 C1720294#C0332285 calificador#atributo

83 nuestro nuestro DP1MSP 164-171 Digestivo220161292_19-04-63 - -

84 Comité comité NCMS000 172-178 Digestivo220161292_19-04-63 - -

85 de de SPS00 179-181 Digestivo220161292_19-04-63 C0332285 atributo

86 Tumores tumor NCMP000 182-189 Digestivo220161292_19-04-63 C0027651 anomalía_morfológica

87 . . Fp 189-190 Digestivo220161292_19-04-63 - -

Unimer

Unimer debe ejecutarse como un módulo al cual le pasamos dos argumentos “input_dir”, con la siguiente estructura:

input_dir/

|_____ doc1.txt

|_____ doc1.txt

|_____ doc1.txt

...

Y el segundo apuntado al directorio de salida, el módulo se ejecutará de la siguiente manera. Primero nos desplazamos a la carpeta unimer:

cd unimer/

Después debemos establecer el entorno de la herramienta:

source venv/bin/activate.csh

Adicionalmente debemos establecer las variables de entorno, si tenemos una terminal Bash ($ detrás de nuestro nombre en la línea de comandos):

source set-env-vars.sh

Y si tenemos una terminal Shell(% detrás de nuestro nombre en la línea de comandos):

source /tartalo01/users/eandres011/set-env-vars-unimer.sh

Y finalmente lanzamos el programa:

python -m unimer path/to/input_dir/ path/to/output_dir/ [--relations]

En este caso podemos decidir si solamente queremos la salida de flairmer + freelingmed, o si adicionalmente [--relations] queremos calcular las relaciones en el texto

El resultado sigue las anotaciones del formato (.conllv2):

output_dir/

|_____ doc1.conll

|_____ doc2.conll

|_____ doc3.conll

…

|_____ output-merged.conll

|_____ output-for-relations.iob

|_____ output-relations.conll

Los atributos que componen la salida son diferentes según el archivo:

1) docX.conll: Contiene para cada fichero la anotación de **flairmer + freelingmed**, sus atributos son los siguientes:

1. **Hitz-zenbakia**: Hitzaren posizioa esaldiarekiko.

2. **Hitz-forma**: Hitzaren testuko forma.

3. **Lema**: Hitzaren lema.

4. **FreePOS**: FreelingMedek eman dion kategoria.

5. **SnoMot**: SnoMedek eman dion mota.

6. **SnoKod**: SnoMedek eman dion kodea.

7. **Offseta**: hitza esaldian ze posiziotan hasten eta bukatzen den adierazten duen tartea.

8. **Dokumentu-izena**: hitza agertzen den dokumentuaren izena.

9. **Esaldi-zenbakia**: hitza dokumentuko zenbatgarren esaldian agertzen den adierazten duen zenbakia.

10. **DeepEnt**: neurona-sare sakonak hitzari esleitu dion entitate motaren etiketa chunkak zehazteko IOB formatuarekin konbinatuta, (B-Grp_Enfermedad, I-Grp_Medicamento, O, B-Alergia).

11. **Ez-Espek**: entitatea eskuzko anotazioan espekulazioa den edo ezeztatuta dagoen adierazten duen etiketa.

2) output-merged.conll: Contiene para cada fichero la anotación de **flairmer + freelingmed + relations** (solamente cuando añadimos el argumento --relations) , sus atributos son los contenidos en docX.conll más los siguientes:

   12. **AH**: Erlazioaren Head ASENT esaldiko zein hitz posizioan dagoen adierazten du.

   13. **ASENT **: Erlazioaren Sentence erlazionatu duen entitatea zein esaldian dagoen adierazten du .

   14. **AERL**: Erlazio mota (Modif, Causada_por eta LOC) adierazten du.

3) output-merged.conll(solamente cuando añadimos el argumento --relations): Contiene para cada fichero la anotación de **flairmer + freelingmed + relations** , sus atributos son los contenidos en docX.conll más los siguientes:

4) output.conll (solamente cuando añadimos el argumento --relations): Es de carácter temporal y en este fichero se compendian todos los textos en un único documento a partir de los ficheros docX.conll.

5) output-for-relations.iob (solamente cuando añadimos el argumento --relations): Es de carácter temporal y contiene todos los pares de relaciones generados a partir de output.conll para el proceso de extracción de relaciones, en este fichero se compendian todos los textos en un único documento.

6) output-relations.conll (solamente cuando añadimos el argumento --relations): Es de carácter temporal y en este fichero se compendian las predicciones del submódulo DeepREL correspondientes a cada línea de output.conll.

De está manera la extracción de relaciones resulta en un fichero **output-merged.conll** que coincide en su totalidad con el expuesto en el formato (.conllv2)

Apéndices

En está sección se expone primeramente la definición del formato (.conllv2) mediante el documento IXAMed CONLL Formatuaren gidalerroak, posteriormente se provee el documento histórico de las actas de reunión.

IXAMed CONLL Formatuaren gidalerroak

Aitziber, Koldo eta Iakes

IXA

2018

Aurkibidea

Formatua

FreelingMed-en eta medikuek eskuzko anotazioren informazioa bildu da esaldika CoNLL formatuaren aldaera batean. Hitz bakoitzaren informazioa 17 zutabetan banatu da:

1. **Hitz zenbakia:** Hitzaren posizioa esaldiarekiko.
2. **Hitz-forma:** Hitzaren testuko forma.
3. **Lema:** Hitzaren lema.
4. **FreePOS:** FreelingMedek eman dion kategoria.
5. **SnoMot:** SnoMedek eman dion mota.
6. **SnoKod:** SnoMedek eman dion kodea.
7. **Offseta:** Hitza esaldian ze posiziotan hasten eta bukatzen den adierazten duen tartea.
8. **Dokumentu izena:** Hitza agertzen den dokumentuaren izena.
9. **Esaldi zenbakia:** Hitza dokumentuko zenbatgarren esaldian agertzen den adierazten duen zenbakia.

Formatua

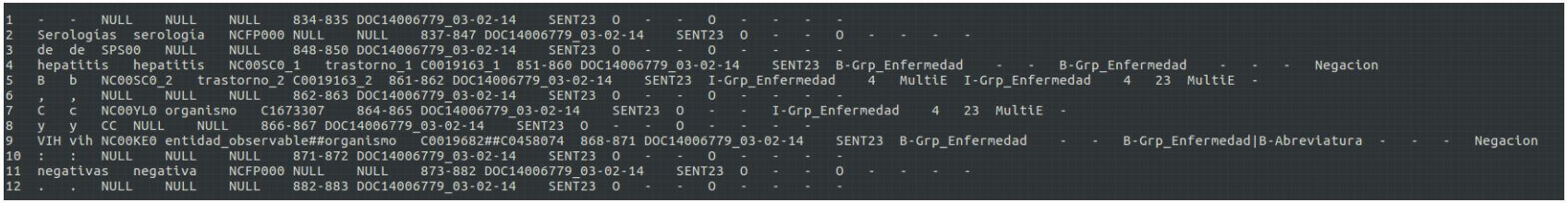
10. **DeepEnt:** Neurona-sare sakonak hitzari esleitu dion entitate motaren etiketa chunkak zehazteko IOB formatuarekin konbinatuta, (B-Grp_Enfermedad, I-Grp_Medicamento, O, B-Alergia).
11. **FreeEnt:** FreelingMedek hitzari esleitu dion entitate motaren etiketa chunkak zehazteko IOB formatuarekin konbinatuta, (B-Grp_Enfermedad, I-Grp_Medicamento, O, B-Alergia).
12. **FreeHead:** FreelingMedek definitutako entitatearen burua esaldiko zein posiziotan dagoen adierazten duen zenbakia.
13. **FreeMultiE:** FreelingMedek definitutako entitateak hitz bat baino gehiago duela adierazteko erabiltzen den etiketa.

Formatua

14. **AnnEnt:** Eskuzko anotazioan hitzari esleitu zaion entitate motaren etiketa chunkak zehazteko erabiltzen den IOB formatuarekin konbinatuta (B-Grp_Enfermedad, I-Grp_Medicamento, O, B-Alergia . . .).
15. **AnnHead:** Eskuzko anotazioan definitutako entitatearen burua esaldiko zein posiziotan dagoen adierazten duen zenbakia.
16. **AnnSent:** Eskuzko anotazioan definitutako entitatearen burua dokumentuko zein esalditan dagoen adierazten duen zenbakia.
17. **AnnErl:** Eskuzko anotazioan definitutako entitatea beste batekin ze erlazioren bidez dagoen lotuta adierazten duen etiketa (Oraingoz MultiE, LOC eta Modif).
18. **Ez/Espek:** Entitatea eskuzko anotazioan espekulazioa den edo ezeztatuta dagoen adierazten duen etiketa. (Etorkizunean gehiago izan daitezke).

Formatua

Adibidea



```
1   -   -   NULL    NULL    NULL    834-835 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
2   Serologías  serología   NCFP000 NULL    NULL    837-847 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
3   de  de  SPS00   NULL    NULL    848-850 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
4   hepatitis   hepatitis   NC00SC0_1   trastorno_1 C0019163_1  851-860 DOC14006779_03-02-14    SENT23  B-Grp_Enfermedad    -   -   B-Grp_Enfermedad    -   -   -   Negacion
5   B   b   NC00SC0_2   trastorno_2 C0019163_2  861-862 DOC14006779_03-02-14    SENT23  I-Grp_Enfermedad    4   MultiE  I-Grp_Enfermedad    4   23  MultiE  -
6   ,   ,   NULL    NULL    NULL    862-863 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
7   C   c   NC00YL0 organismo   C1673307    864-865 DOC14006779_03-02-14    SENT23  O   -   -   I-Grp_Enfermedad    4   23  MultiE  -
8   y   y   CC  NULL    NULL    866-867 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
9   VIH vih NC00KE0 entidad_observable##organismo   C0019682##C0458074  868-871 DOC14006779_03-02-14    SENT23  B-Grp_Enfermedad    -   -   B-Grp_Enfermedad|B-Abreviatura  -   -   -   Negacion
10  :   :   NULL    NULL    NULL    871-872 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
11  negativas   negativa   NCFP000 NULL    NULL    873-882 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
12  .   .   NULL    NULL    NULL    882-883 DOC14006779_03-02-14    SENT23  O   -   -   O   -   -   -   -
```

HZ	HF	FE	FH	FM	AE	AH	AERL	EZ/ESPEK
1	Serologías	O	-	-	O	-	-	-
2	de	O	-	-	O	-	-	-
3	hepatitis	B-Grp_Enf	-	-	B-Grp_Enf	-	-	Negacion
4	B	I-Grp_Enf	3	MultiE	I-Grp_Enf	3	MultiE	-
5	,	O	-	-	O	-	-	-
6	C	O	-	-	I-Grp_Enf	3	MultiE	-
7	y	O	-	-	O	-	-	-
8	VIH	B-Grp_Enf	-	-	B-Grp_Enf	-	-	Negacion
9	negativas	O	-	-	O	-	-	-

Hitzen banaketa

Eskuzko anotazioa eta FreelingMedekoa corpus berean biltzeko hitz-anitzeko entitateak banatu egin dira. Ondorengo adibidean entitate osoaren NC00SC0, anomalia_morfologica eta C2826025 ezaugarriak entitatea osatzen duten hitz bakoitzean errepikatzen dira:

HZ	HF	POS	SM	SK	FE	FH	FM
1	Leucemia	NC00SC0_1	anomalía_morfológica	C2826025	B-Grp_Enf	-	-
2	aguda	NC00SC0_2	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE
3	de	NC00SC0_3	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE
4	Fenotipo	NC00SC0_4	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE
5	Mixto	NC00SC0_5	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE

Erlazioak

Zeintzuk erlazio kodetu dira?

- Modif, Causada_por eta LOC.
- etorkizunean ann fitxategietan besteren bat sartzen bada, bihurketarako programa prestatuta dago horiek ere harrapatzeko.
- entitate batek hitz bat baino gehiago duenean MultiE erlazioa erabili dugu.

Erlazioak

Nola kodetu dira erlazioak?

- Eskuzko anotazioan dagoen erlazio bat kodetzeko mendeko hitza zein erlazio bidez lotuta dagoen adierazten da azken zutabean (AnnErl eremua).
- Hitzak definitutako erlazioa zenbatgarren esaldiko zenbatgarren hitzarekin duen ere esan behar da AnnHead eta AnnSent eremuekin.

HZ	HF	POS	SM	SENT	AE	AH	ASENT	AERL
7	Megalomaniaca	AQ0FS0	NULL	SENT42	B-Calificador	1	42	**Modif**

Erlazioak

Hitz-anitzeko entitateak nola kodetu dira?

- Eskuzko anotazioan entitate batek hitz bat baino gehiago duenean, entitate horretako hitz guztiak entitate horretako lehenengo hitzera lotu dira MultiE erlazioaren bidez.

HZ	HF	POS	SM	SK	FE	FH	FM
1	Leucemia	NC00SC0_1	anomalía_morfológica	C2826025	B-Grp_Enf	-	-
2	aguda	NC00SC0_2	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE
3	de	NC00SC0_3	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE
4	Fenotipo	NC00SC0_4	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE
5	Mixto	NC00SC0_5	anomalia_morfologica	C2826025	I-Grp_Enf	1	MultiE

Erlazioak

Hitz-anitzeko entitateak nola kodetu dira?

- Hitz bat baino gehiagoko entitate bat beste entitate batekin erlazionatzeko entitateren lehenengo hitzari esleitzen zaio erlazio etiketa.

HZ	HF	POS	SM	SENT	AE	AH	ASENT	AERL
1	Insomnio	NC00SC0_1	trastorno	SENT45	B-Grp_Enf	-	-	-
2	de	NC00SC0_2	trastorno	SENT45	B-Calificador	1	45	Modif
3	conciliación	NC00SC0_3	trastorno	SENT45	I-Calificador	2	45	MultiE

Erlazioak

Hitz-anitzeko entitate ez jarraituak nola kodetu dira?

- Eskuzko anotazioan hitz-anitzeko entitate ez jarraituren bat baldin badago jarraituetan egiten den gauza bera egin da, hots, hitz guztiak lehenengo hitzera lotu MultiE etiketaren bidez.
- Hurrengo adibidean *C* hitza *hepatitis*ekin lotu da MultiE etiketaren bidez *hepatitis C* entitatea definitzeko.

HZ	HF	FE	FH	FM	AE	AH	AERL	EZ/ESPEK
1	Serologías	O	-	-	O	-	-	-
2	de	O	-	-	O	-	-	-
3	hepatitis	B-Grp_Enf	-	-	B-Grp_Enf	-	-	Negacion
4	B	I-Grp_Enf	3	MultiE	I-Grp_Enf	3	MultiE	-
5	,	O	-	-	O	-	-	-
6	C	O	-	-	I-Grp_Enf	3	MultiE	-
7	y	O	-	-	O	-	-	-
8	VIH	B-Grp_Enf	-	-	B-Grp_Enf	-	-	Negacion
9	negativas	O	-	-	O	-	-	-

Erlazioak

Entitate bat entitate bat baino gehiagorekin erlazionatuta dagoenean nola kodetu da?

- AnnHead, AnnSent eta AnnErl eremuetan zehazten dira erlazio desberdinak eta zein entitaterekin duten erlazioa. Adibidez:

HZ	HF	SENT	AE	AH	ASENT	AERL
8	VIH	SENT15	B-Grp_Enf	3,15	15,14	Causada_por,Modif

Ezeztapena/Espekulazioa

- Ezeztapena edo espekulazioa adierazteko Ez/Espek eremua erabili dugu. Eskuzko anotazioan hitz bat ezeztatuta dagoela edo espekulazioa dela adierazten bada Ez/Espek zutabean Negacion edo Especulacion etiketa idatzi behar da, hurrenez hurren. Adibidez:

HZ	HF	SENT	AE	AH	ASENT	AERL	EZ/ESPEK
1	Sin	SENT10	O	-	-	-	-
2	cancer	SENT10	B-Grp_Enf	-	-	-	Negacion
3	renal	SENT10	I-Grp_Enf	2	10	LOC	-
4	y	SENT10	O	-	-	-	-
5	probable	SENT10	O	-	-	-	-
6	linfoma	SENT10	B-Grp_Enf	-	-	-	Especulacion

Entitate motak

- Corpusean erabili diren entitate motak eskuzko anotazion erabili eta FreelingMedek sortzen dituen berberak dira, hau da, Grp_Enfermedad, Grp_Medicamento, Alergia, Abreviatura . . .
- Bestalde, etiketa horiei IOB etiketak gehituko dizkiegu hasieran.

Entitate motak

Nola kodetu dira hitz bakarreko entitateak?

- Hitz bakarreko entitateei B (begin) IOB etiketa esleitzen zaie aurretik. Adibidez:

HZ	HF	SENT	AE	AH	ASENT	AERL	EZ/ESPEK
6	linfoma	SENT10	B-Grp_Enf	-	-	-	Especulacion

Entitate motak

Nola kodetu dira hitz-anitzeko entitateak?

- Hitz-anitzeko entitatearen lehenengo hitzari B (begin) IOB etiketa esleitzen zaio aurretik eta ondorengoei I (in) etiketa. Adibidez:

HZ	HF	SENT	AE	AH	ASENT	AERL	EZ/ESPEK
1	Sin	SENT10	O	-	-	-	-
2	cancer	SENT10	B-Grp_Enf	-	-	-	Negacion
3	renal	SENT10	I-Grp_Enf	2	10	LOC	-

Entitate motak

Hitz-anitzeko entitate ez jarraituak nola kodetu dira?

- Jarraituak diren entitateak bezala. Hitz-anitzeko entitatearen lehenengo hitzari B (begin) IOB etiketa esleitzen zaio aurretik eta ondorengoei I (in) etiketa.
- Ondorengo adibidean *hepatitis C* entitate ez jarraituko *hepatitis* hitza B-Grp_Enfermedad bezala etiketatzen da eta *C* hitza I-Grp_Enfermedad bezala nahiz eta tartean O (Out) bat egon.

HZ	HF	FE	FH	FM	AE	AH	AERL	EZ/ESPEK
1	Serologías	O	-	-	O	-	-	-
2	de	O	-	-	O	-	-	-
3	hepatitis	B-Grp_Enf	-	-	B-Grp_Enf	-	-	Negacion
4	B	I-Grp_Enf	3	MultiE	I-Grp_Enf	3	MultiE	-
5	,	O	-	-	O	-	-	-
6	C	O	-	-	I-Grp_Enf	3	MultiE	-
7	y	O	-	-	O	-	-	-
8	VIH	B-Grp_Enf	-	-	B-Grp_Enf	-	-	Negacion
9	negativas	O	-	-	O	-	-	-

Entitate motak

Solapatzen diren hitz-anitzeko entitateak nola kodetu dira?

- Zenbait kasutan eskuzko anotazioan hitz berdina bi entitate desberdinetan ager daiteke. Adibidez:
  - T17 Grp_Enfermedad 1442 1462 Doble lesión **aórtica**
  - T18 Grp_Enfermedad 1455 1462;1481 1508 **aórtica** regurgitación leve-moderada
- *aórtica* hitza bi entitate desberdinetan agertzen da, batean hasierako hitza da eta bestean bukaerakoa. Bi etiketa jarri dizkiogu, I-Grp_Enfermedad eta B-Grp_Enfermedad.

HZ	HF	AE	AH	AERL
1	Doble	B-Grp_Enf	-	-
2	lesión	I-Grp_Enf	1	MultiE
3	aórtica	I-Grp_Enf|B-Grp_Enf	1	MultiE
4	:	O	-	-
5	estenosis	B-Grp_Enf	-	-
6	leve	I-Grp_Enf	5	Modif
7	y	O	-	-
8	regurgitación	I-Grp_Enf	3	MultiE
9	leve-moderada	I-Grp_Enf	3	MultiE

Kasu bereziak

- Fitxategietan ikusi ditugun kasuen artean batzuk bereziak izan dira. Gure asmoa kasu berezi horiek gidalerro honetan zehaztea da kasu horiek tratatzeko jarraitu ditugun irizpideekin batera.

Kasu bereziak

Kasu bereziak 1

- Entitate bat osatzen duten hitzen offsetak bat ez datozenean eskuzko anotazioan eta FreelingMeden irteeran.Adibidez:
  - **Eskuzkoan**:
  - T9 Grp_Enfermedad 1066 1109 Leucemia aguda de Fenotipo Mixto T/mieloide
  - **FreelingMed:**
  - Leucemia_aguda_de_Fenotipo_Mixto
  - T
  - /
  - mieloide
- Eskuzko anotazioan *T/mieloide* dena batera dago idatzita eta CoNLL-an banatuta, orduan offsetak desberdinak dira.

Kasu bereziak

Kasu bereziak 1 / Tratamendurako irizpideak 1

- *T / mieloide* osatzen duten elementu bakoitza *Leucemia aguda de Fenotipo Mixto T/mieloide* entitatearen hitz bat izango balitz bezala tratatu da.

HZ	HF	AE	AH	AERL
1	Leucemia	B-Grp_Enf	-	-
2	aguda	I-Grp_Enf	1	MultiE
3	de	I-Grp_Enf	1	MultiE
4	Fenotipo	I-Grp_Enf	1	MultiE
5	Mixto	I-Grp_Enf	1	MultiE
6	T	I-Grp_Enf	1	MultiE
7	/	I-Grp_Enf	1	MultiE
8	mieloide	I-Grp_Enf	1	MultiE

Kasu bereziak

Kasu bereziak 1 / Tratamendurako irizpideak 2

- *T / mieloide* elementua hitz-anitzeko entitate bat balitz bezala hartu eta gero horren burua *Leucemia aguda de Fenotipo Mixto T/mieloide* entitatearen buruarekin lotu da.

HZ	HF	AE	AH	AERL
1	Leucemia	B-Grp_Enf	-	-
2	aguda	I-Grp_Enf	1	MultiE
3	de	I-Grp_Enf	1	MultiE
4	Fenotipo	I-Grp_Enf	1	MultiE
5	Mixto	I-Grp_Enf	1	MultiE
6	T	I-Grp_Enf	1	MultiE
7	/	I-Grp_Enf	6	MultiE
8	mieloide	I-Grp_Enf	6	MultiE

Eranskina

- **Corpusari buruzko zehaztasunak:**
- Miriamek eta Sarak etiketatu dituzten fitxategi guztiak bihurtu dira:
  - ixamed/Corpusa/Galdakao/EtiquetadosMiriam
  - ixamed/Corpusa/Basurto/EtiquetadosSara
- **Corpusaren helbidea:**
- ixamed/Corpusa/Berdeen_Corpusa/

Eranskina

- **Bihurketan erabilitako tresnak:**
- ixamed/Corpusa/Berdeen_Corpusa/Erabilitako_Tresnak
- **Nola exekutatu:**
- ann-ak eta txt fitxategiak dauden katalogoan sartu tresnak eta *Conllv2formatuan_lortu_informazioa.sh* scripta egikaritu.

Eranskina

- **Gidalerro honetan edo bihurketa egiteko tresnetan aldaketak sartu edo hobetuz gero mesedez jakinarazi:**
  - Iakes Goenaga: iakesg@gmail.com
  - Koldo Gojenola: koldo.gojenola@ehu.eus
  - Aitziber Atutxa: aitziber.atucha@ehu.eus