Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA

Datos en Salud para tecnologías del lenguaje

Martin Krallinger, Siamak Barzegar

(BSC-CNS)

mkrallin@bsc.es

Ejes fundamentales para promover el desarollo de TL aplicado al dominio salud

Barreras y potenciales soluciones para fomentar el desarrollo de PLN clínico

Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions FREE

Wendy W Chapman ✉, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, Ozlem Uzuner

Journal of the American Medical Informatics Association, Volume 18, Issue 5, September 2011, Pages 540–543, https://doi.org/10.1136/amiajnl-2011-000465

Published: 01 September 2011 **Article history** ▾

Névéol *et al. Journal of Biomedical Semantics* (2018) 9:12
https://doi.org/10.1186/s13326-018-0179-8

Journal of Biomedical Semantics

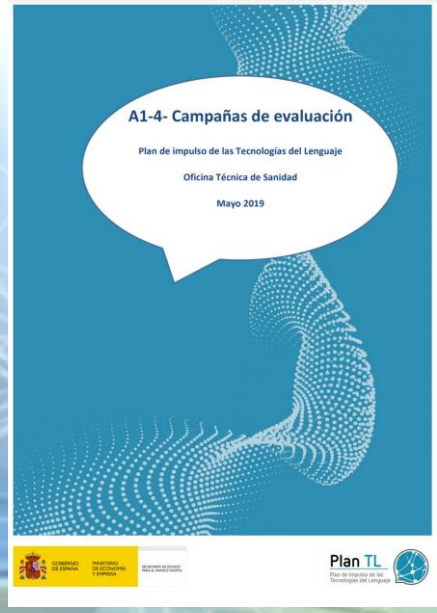
REVIEW Open Access

Clinical Natural Language Processing in languages other than English: opportunities and challenges

Aurélie Névéol[1]*, Hercules Dalianis[2], Sumithra Velupillai[3,4], Guergana Savova[5] and Pierre Zweigenbaum[1]

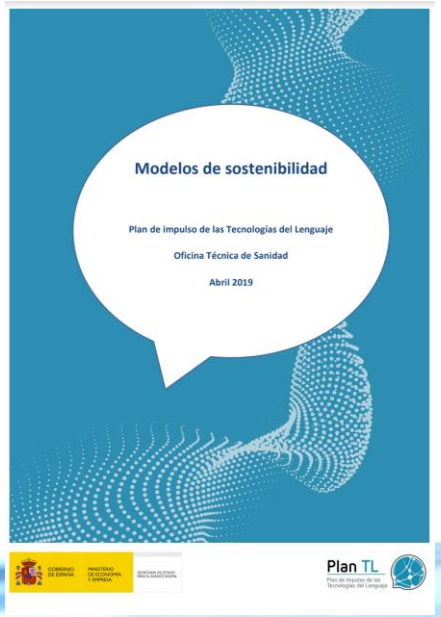
Estudios previos del Plan TL

Campañas de evaluación

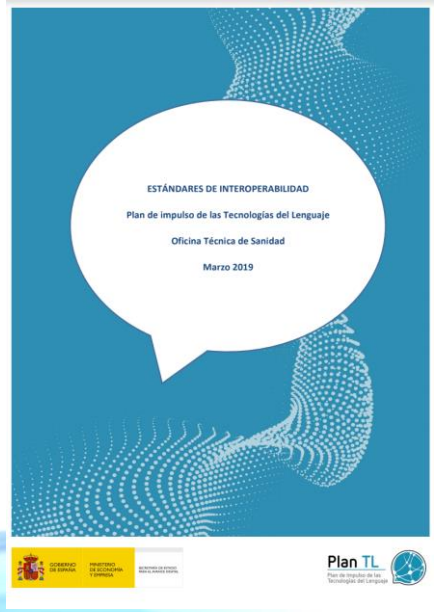


Aspectos legales

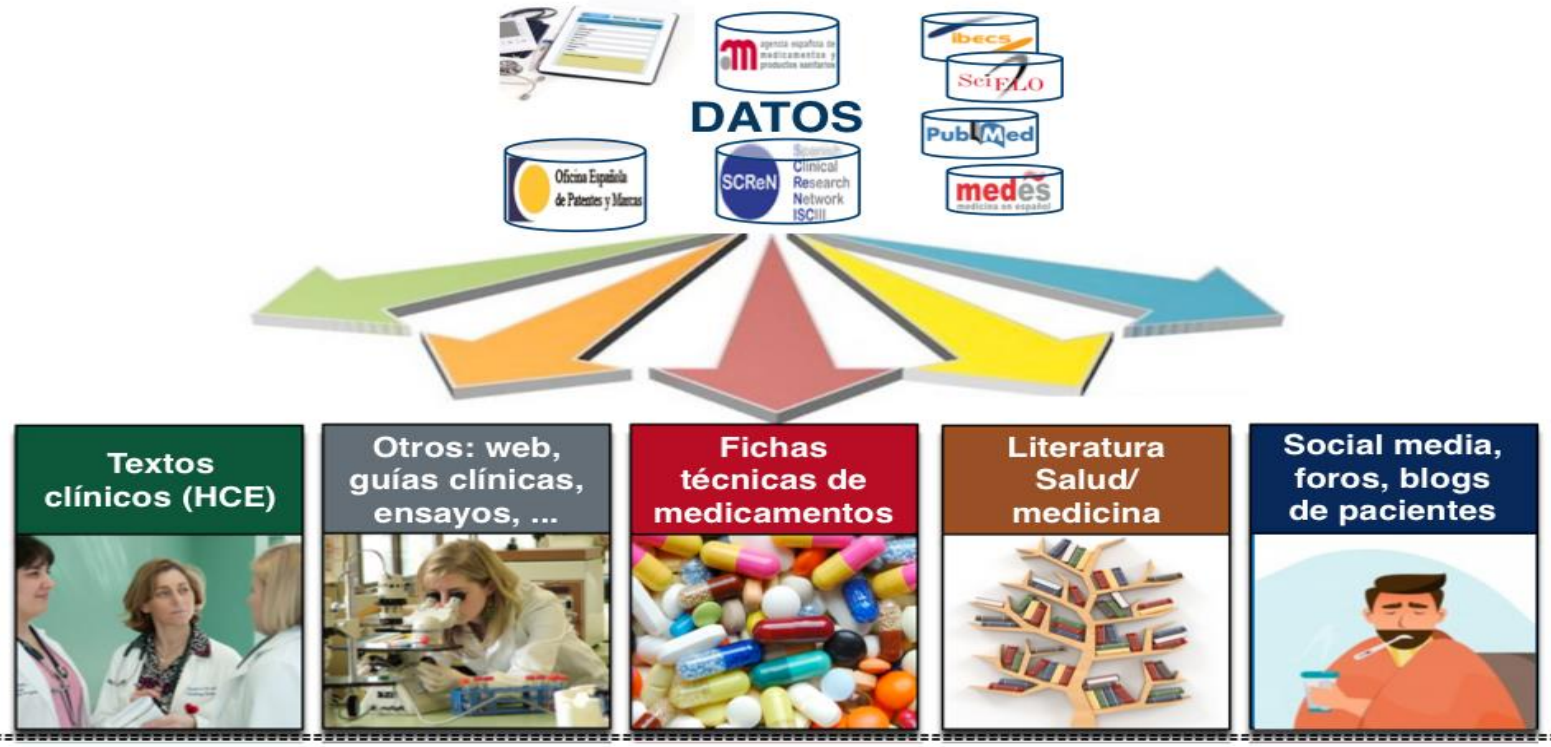
Sostenibilidad de recursos



Interoperabilidad (formatos)



Tipos de datos/documentos/anotaciones de relevancia para tecnologías del lenguaje



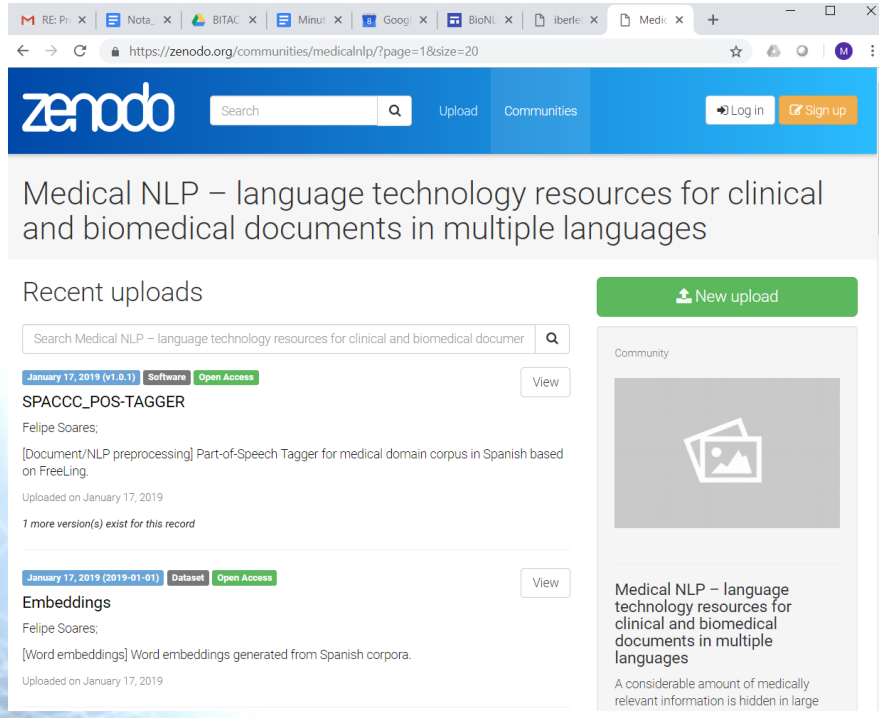
Tipo de anotación/etiquetas

Datos sin anotación	Documentos clasificados/ indizados	Textos etiquetados (text-bound annotations)	Corpus paralelos o comparables
* Modelos del lenguaje * Terminologias * Distant supervision * Annotaciones Silver Standard/automáticas	* Clasificación automatica * Recuperación de información * Indización * Codificación clínica * Sistemas pregunta/ respuesta	* Anotación semántica * Reconocimiento de conceptos, entidades,.. * Reconocimiento de eventos/relaciones,... * Knowledge discovery *Otras	* Traducción automatica * Extraction de glosarios y terminologias bilingues

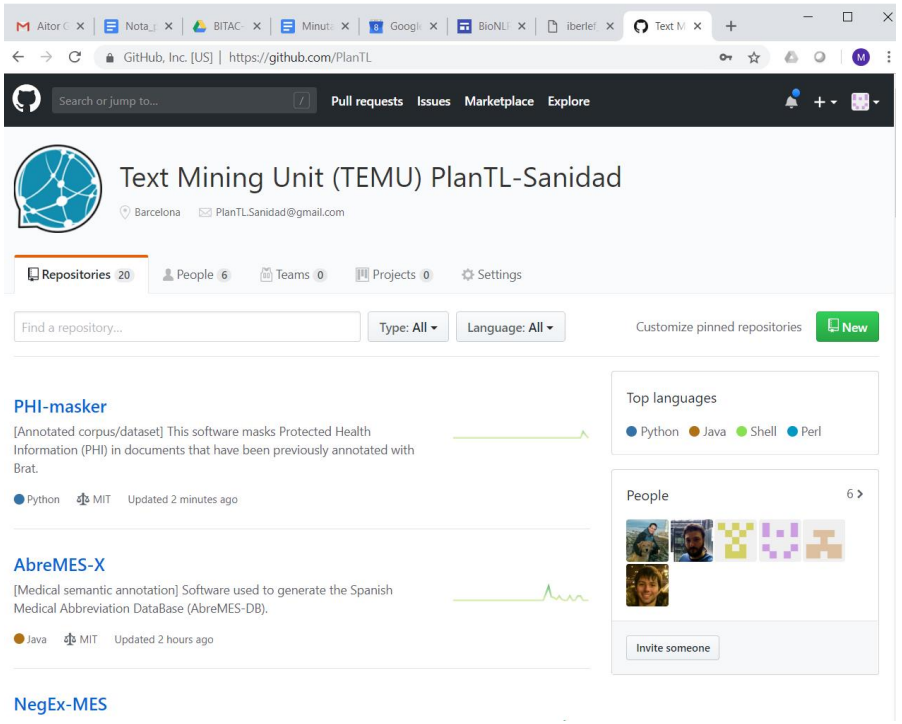
Repositorios abiertos de recursos (FAIR)

zenodo

GitHub



https://zenodo.org/communities/medicalnlp



https://github.com/PlanTL

Datos dominio salud de relevancia para TL

July 23, 2019 (v1) | Dataset | Open Access

View

BVS Corpus: A Multilingual Parallel Corpus and Translation Experiments of Biomedical Scientific Texts

Felipe Soares; Martin Krallinger;

The BVS database (Health Virtual Library) is a centralized source of biomedical information for Latin America and Carib, created in 1998 and coordinated by BIREME in agreement with the Pan American Health Organization (OPAS). Abstracts are available in English, Spanish, and Portuguese, with a s

Uploaded on July 23, 2019

➢ Corpus para desarrollo de sistemas de traducción automática

July 15, 2019 (v2) | Dataset | Open Access

View

LILACS and IBECS annotated abstracts in Spanish

PlanTL - Sanidad;

http://temu.bsc.es/mesinesp/

Annotated articles from IBECS and LILACS, where annotated means that MeSH/DeCS terms have been assigned to the articles by the human curators in IBECS and LILACS.

Uploaded on July 15, 2019

➢ Corpus para desarrollo de sistemas de indización automática

1 more version(s) exist for this record

https://zenodo.org/communities/medicalnlp

Corpus para desarrollo de sistemas de indización automática: tarea MESINESP/BioASQ


```
{
    "journal": "Rev. cientif. cienc. med",
    "title": "Intervención quirúrgica de linfedema escrotal gigante, bolivia",
    "db": "LILACS",
    "id": "biblio-1003801",
    "decsCodes": [
        "21044",
        "9562",
        "12992",
        "8386",
        "21034",
        "331",
        "13883"
    ],
    "year": 2019,
    "abstractText": "Linfedema escrotal es una patología de escasa frecuencia como presentación idiopática. Se conoce también como elefantiasis por las modificaciones que produce en tejido dérmico, se clasifica de acuerdo a la edad de aparición en congénito o adquirido. El diagnóstico es clínico y la etiología se confirma con exámenes complementarios. El tratamiento recomendado es quirúrgico aunque se puede recurrir a tratamientos paliativos en casos de menor gravedad. Se presenta paciente procedente de Cochabamba-Bolivia, masculino de 33 años con cuadro de 3 años de evolución, se realizó la extirpación de 3,7 Kg de tejido escrotal linfedematoso que tras la intervención quirúrgica presentó una evolución favorable sin complicaciones, mejorando la calidad de vida del paciente y el cuadro clinico."
}
```


Un total de 318,658 entradas con al menos un concepto (código DeCS)

http://temu.bsc.es/mesinesp/

MeSpEn colección de recursos para TL salud

MeSpEn: the resource for English-Spanish Medical Machine Translation and Terminologies:

Census of Parallel Corpora, Glossaries and Term Translations

Biomedical and clinical literature | Patient information: MedlinePlus | Glossaries | External links

- Recurso que integra y armoniza (formato Dublin core y TEI):
- IBECS: 168,198 resúmenes de publicaciones médicas
- SCIELO: 161,710 Free Open Access (OAI-PMH)
- PubMed: 127,61 resúmenes en de publicaciones médicas español, y 300,690 títulos
- Contenido Web Salud: 7,033 articles en MedlinePus

http://temu.bsc.es/mespen

Villegas M, et al. The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. In Proceedings of the LREC 2018 Workshop. MultilingualBIO: Multilingual Biomedical Text Processing, 32-39, 2018.

Corpus Web Salud Español (CoWeSE)

- Datos crudos corpus web / crawling > de medio tera (en proceso)

- Semilla de crawler de más de 4500 dominio web seleccionados manualmente (incluye LIS – Lugares de Interés de Salud, ISCIII).

- Datos en español, catalán, gallego y euskera

- Diversidad de tipos de contenido: (1) sociedades médicas, (2) sociedades científicas, (3) revistas electrónicas abiertas, (3) centros de investigación, (4) empresas farmacéuticas, (5) webs de educativas y divulgación de salud, (6) asociaciones de pacientes, (7) blogs y páginas personales de profesionales de la salud, (8) webs de centros asistenciales e hospitales (públicos y privados), (9) colegios profesionales, (10) páginas de instituciones y organizaciones de salud publica.

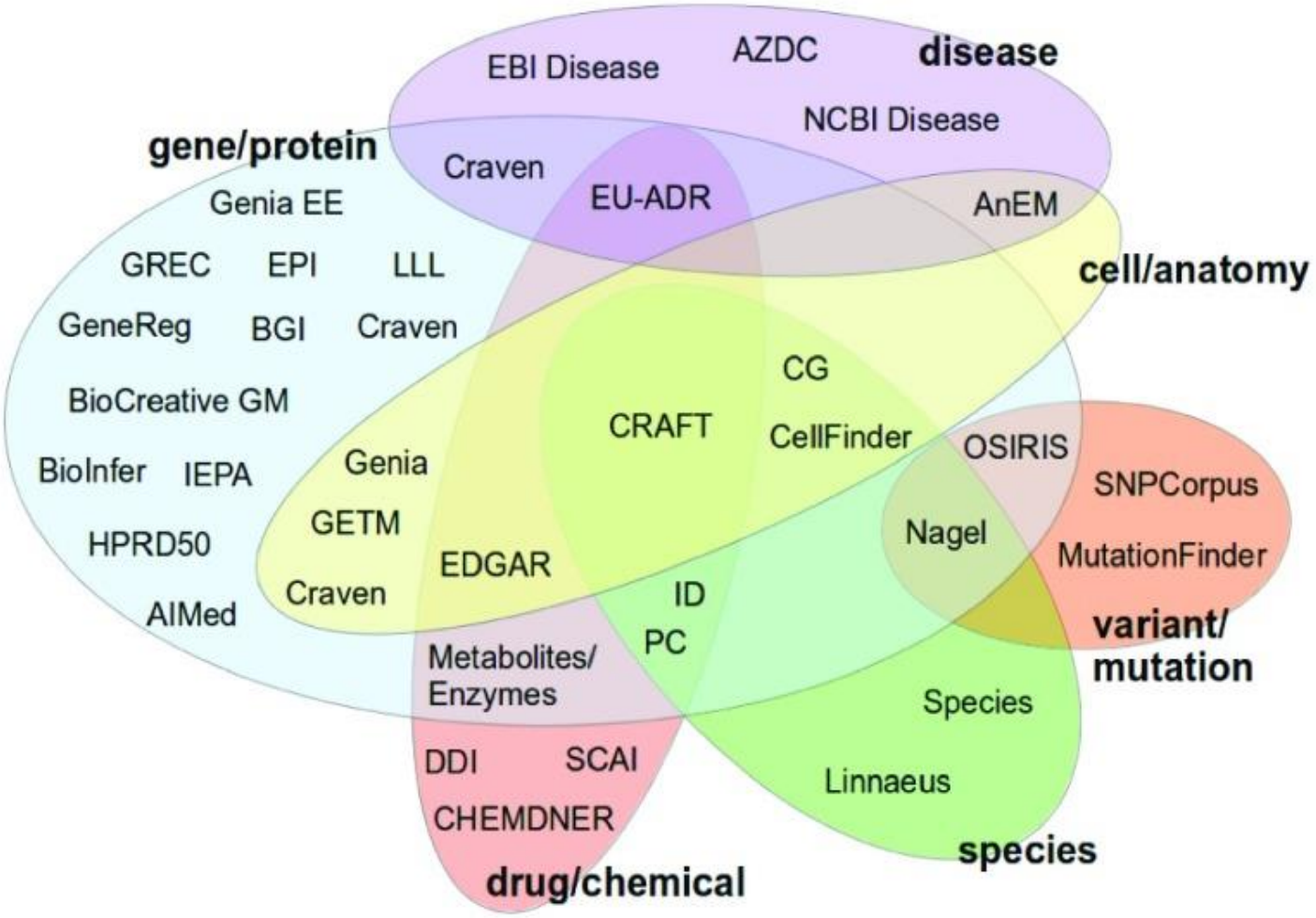
- Uso: modelos del lenguaje, terminologías, corpus sintéticos salud (Principio de Web as a corpus, por ejemplo Common Crawl: 215 TB).

Corpus anotados para sistemas PLN en salud

- Colección de documentos que generalmente pertenece a un tema en particular y que se ha anotado de acuerdo con un esquema predefinido.
- Recurso clave para desarrollar nuevos métodos en la minería de textos basados en inteligencia artificial (aprendizaje de maquina supervisado)
- Permiten la comparación entre sistemas objetiva e independiente.
- Permiten la reproducibilidad de experimentos.
- Tipicamente son ‘text-bound annotations’ o corpus (etiquetado de menciones y relaciones en el texto)

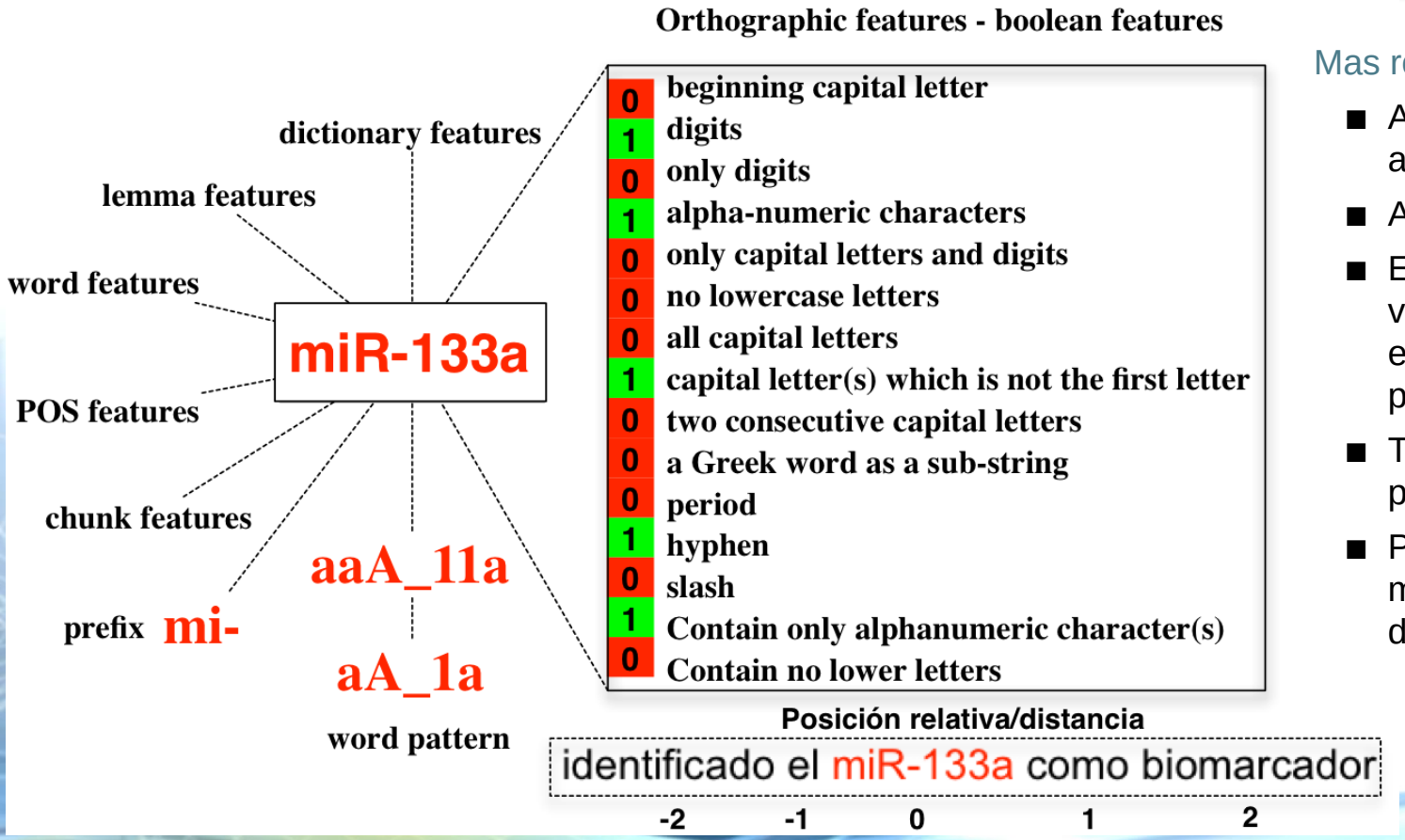
Corpus anotados (textos etiquetados) en biomedicina (inglés)

- Corpus anotados BioNLP (Fuente: Mariana Neves. *An analysis on the entity annotations in biological corpora*)



Corpus y textos etiquetados para sistemas de IA: ejemplo ilustrativo

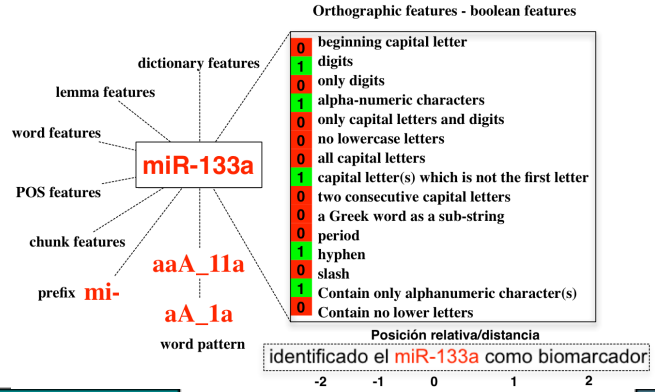
Nuestro estudio es el primero que ha identificado el miR-133a como biomarcador del cáncer de pulmón. Su función es rezeducir la FOXQ1 e inhibir la transición epitelio-mesenquimatosa, la cual antagoniza la génesis tumoral en el cáncer de pulmón. Por consiguiente, nuestros datos respaldan el papel del miR-133a como posible instrumento terapéutico molecular en el tratamiento del cáncer de pulmón.

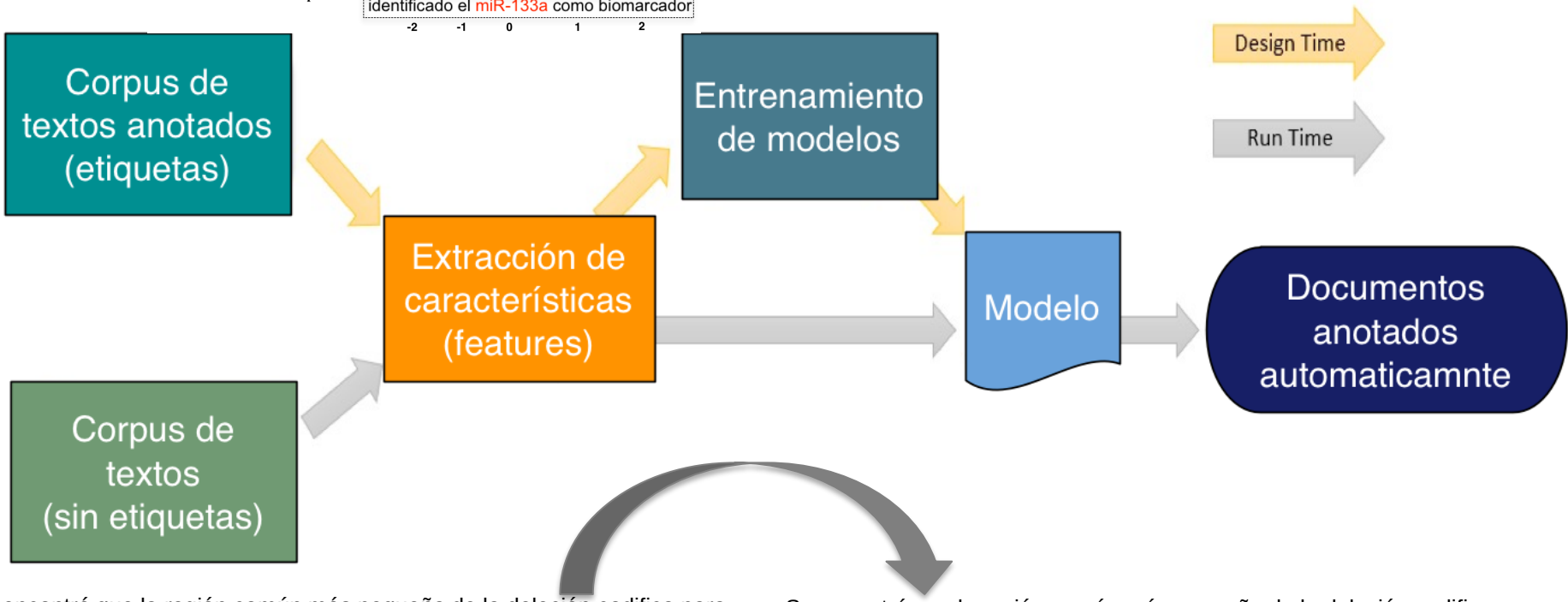


Mas robusto para:

- Acrónimos y abreviaturas.
- Ambigüedad.
- Errores ortográficos, variantes tipográficas, errores de escritura, puntuación, acentuación
- Términos médicos multi-palabra y variabilidad.
- Polisemia (palabra con mas de un sentido) y desambiguación.

Corpus para generación de modelos y entrenamiento de sistemas PLN basados en IA

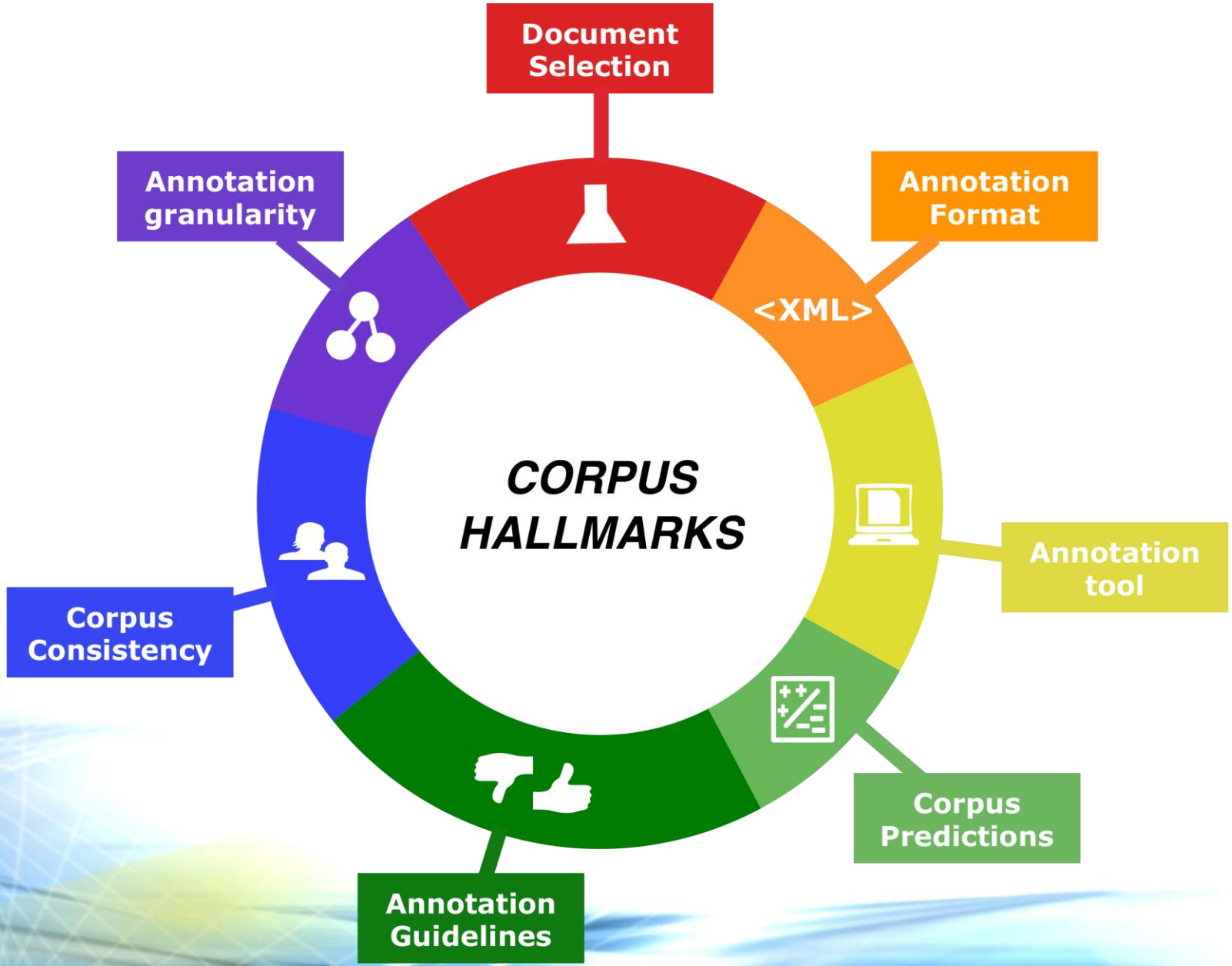




... Se encontró que la región común más pequeña de la deleción codifica para 2 miRNAs: miR-15a y miR-16-1, lo que sugiere su papel como genes supresores de tumores. La ausencia de miR-15a y miR-16-1 induce niveles elevados de esta proteína y el bloqueo de la apoptosis.30,31 Otros ejemplos de miRNAs que funcionan como supresores tumorales son las familias de miR-34 y ....

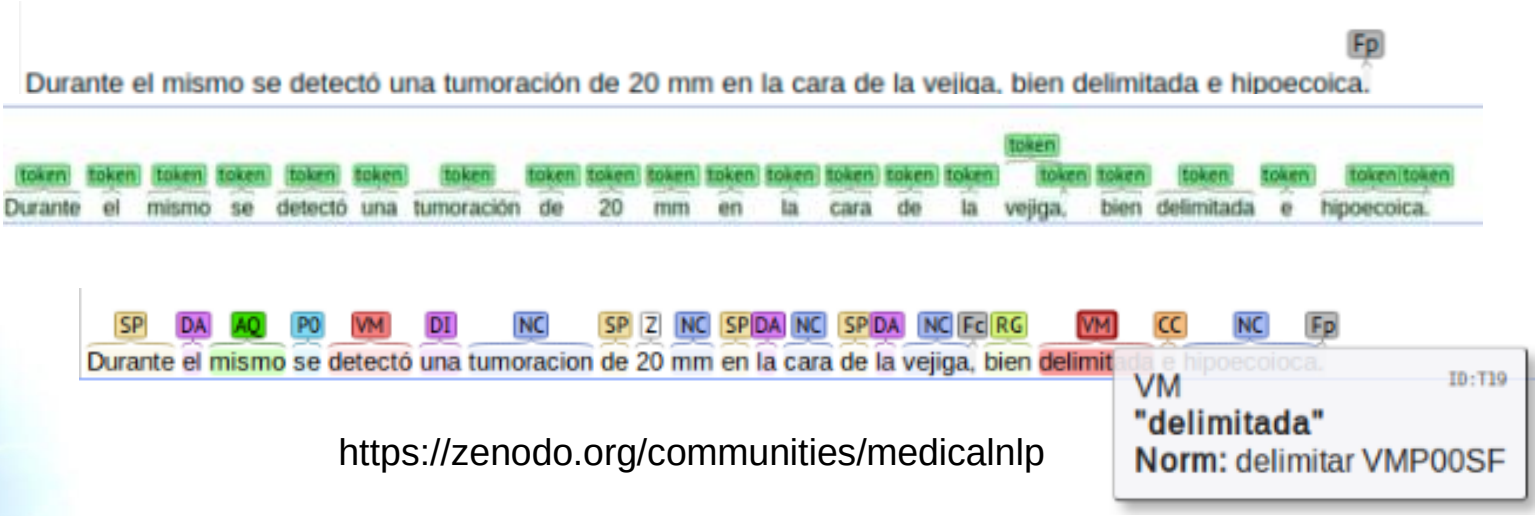
... Se encontró que la región común más pequeña de la deleción codifica para 2 miRNAs: miR-15a y miR-16-1, lo que sugiere su papel como genes supresores de tumores. La ausencia de miR-15a y miR-16-1 induce niveles elevados de esta proteína y el bloqueo de la apoptosis.30,31 Otros ejemplos de miRNAs que funcionan como supresores tumorales son las familias de miR-34 y ....

Criterios relevantes para construccion de corpus anotados del Plan TL Salud



Corpus Gold Standard Plan TL: pre-procesamiento y anotación morfosintáctica

- **SPACCC_POS**
  -Anotación de marcas de límite de oración, tokenización, lema e información categoría morfo-sintáctica.
  -Guía/esquema de anotación y análisis de calidad (consistencia)
  -Uso para adaptación al dominio medico de FreeLing

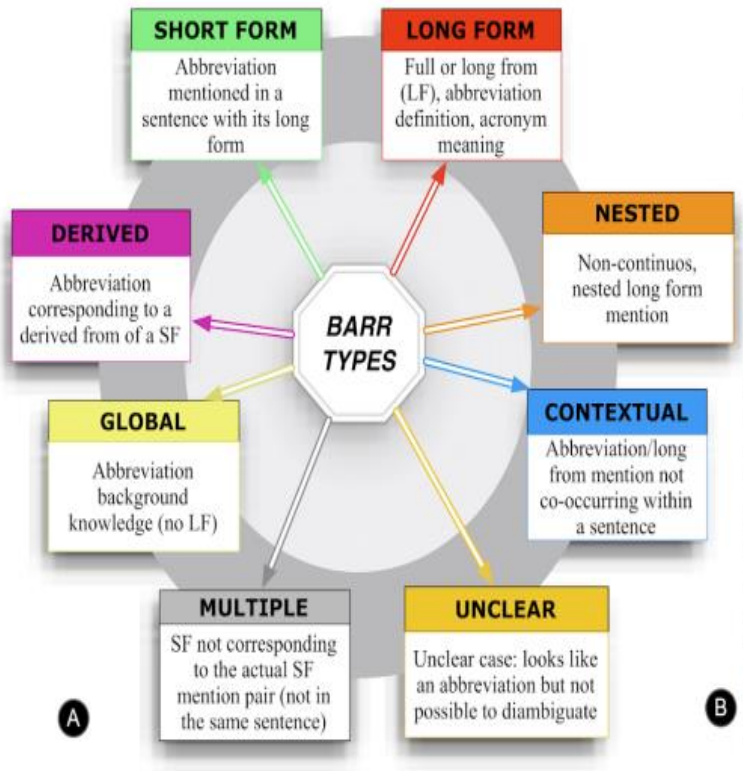


https://zenodo.org/communities/medicalnlp

SPACCC POS Tagger Spanish Clinical Case Corpus Part-of-Speech Tagger	http://temu.bsc.es/demos/spaccc-pos-tagger

Creación de corpus Gold Standard anotados: abreviaturas, siglas y acrónimos

- **Corpus BARR y BARR2 (campaña de evaluación IberEval)**
  Anotaciones de abreviaturas y sus correspondientes formas largas en títulos y resúmenes de artículos médicos y casos clínicos.



"Enfermedad tromboembólica venosa idiopática versus secundaria. Hallazgos del registro RIETE"

12204

"Antecedentes y objetivos. El Registro Informatizado de Enfermedad Tromboembólica (RIETE) es un registro prospectivo que incluye de forma consecutiva pacientes diagnosticados de enfermedad tromboembólica venosa. Hemos comparado la presentación clínica y la respuesta al tratamiento anticoagulante en pacientes con enfermedad tromboembólica venosa idiopática (ETEVI) versus secundaria (ETEVS, asociada a algún factor de riesgo). Pacientes y métodos. Se analizaron las diferencias en las características clínicas, comorbilidad, tratamiento y episodios durante los primeros 3meses tras el diagnóstico de ETEV en los pacientes con ETEVI o ETEVS y según su presentación clínica inicial. Resultados. Se incluyeron 39.921 pacientes, con ETEVI (n=18.029; 45,1%) o ETEVS (n=21.892; 54,9%). Los pacientes con ETEVI mostraron más antecedentes de ETEV que los diagnosticados de ETEVS (p<0,001). El tratamiento inicial fue similar en ambos grupos, pero se colocaron más filtros de vena cava inferior en el grupo de ETEVS (p<0,001). A largo plazo se utilizó con mayor frecuencia heparina de bajo peso molecular en el grupo de ETEVS que en el de ETEVI. A los 90días, la recidiva de ETEV, el sangrado y la muerte fueron significativamente más frecuentes en el grupo con ETEVS. El análisis multivariante confirmó que la ETEVI se asoció a un menor número de sangrados mayores (OR, 0,60; IC95%, 0,50-0,61; p<0,001) y mortales (OR, 0,41; IC95%, 0,29-0,62; p<0,001), menor número de recidivas (OR, 0,58; IC95%, 0,39-0,78; p<0,001) y de embolismo pulmonar mortal (OR, 0,29; IC95%, 0,12-0,52; p<0,001). Estas diferencias se mantuvieron en los pacientes cuya ETEV se inició con un embolismo pulmonar o con una trombosis venosa profunda. Conclusiones. La ETEVI tiene mejor pronóstico que la ETEVS a los 90días del diagnóstico (AU)"

B

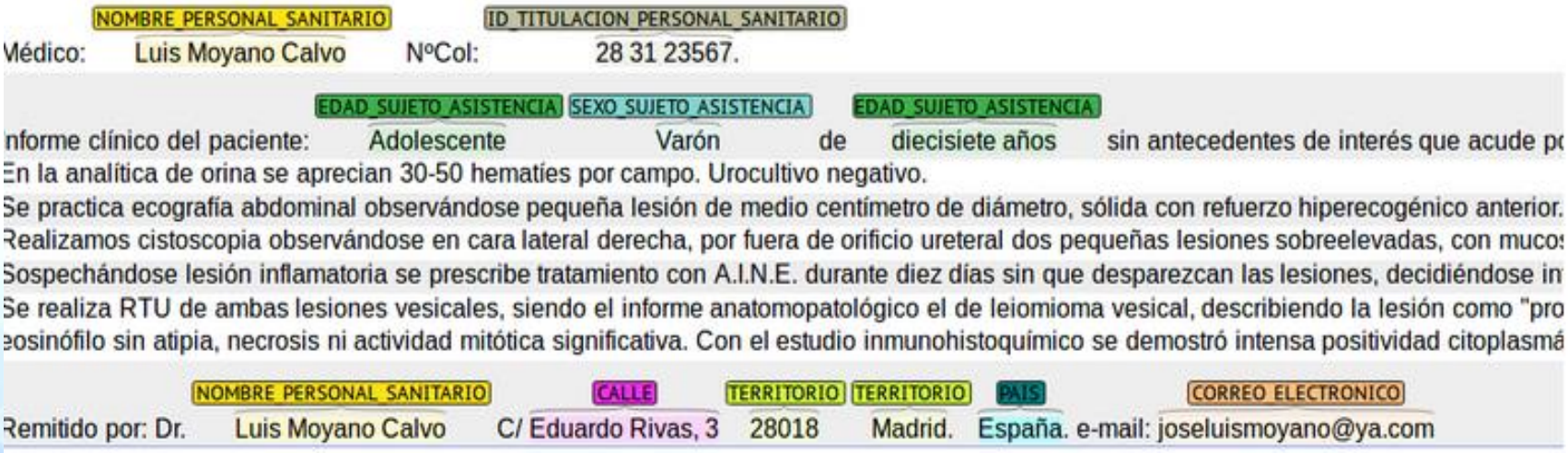
http://temu.bsc.es/BARR

http://temu.bsc.es/BARR2

Creación de corpus Gold Standard de entidades de carácter personal

Corpus anotados por el PlanTL para tarea de anonimizacion y de-identificacion: campaña de evaluacion MEDDOCAN (IberLEF)

- **Corpus MEDDOCAN**
  - Anotación de información de salud protegida.
  - Guía/esquema de anotación y análisis de calidad (consistencia)



http://temu.bsc.es/meddocan/

Creación de corpus Gold Standard para reconocimiento de entidades de carácter personal

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

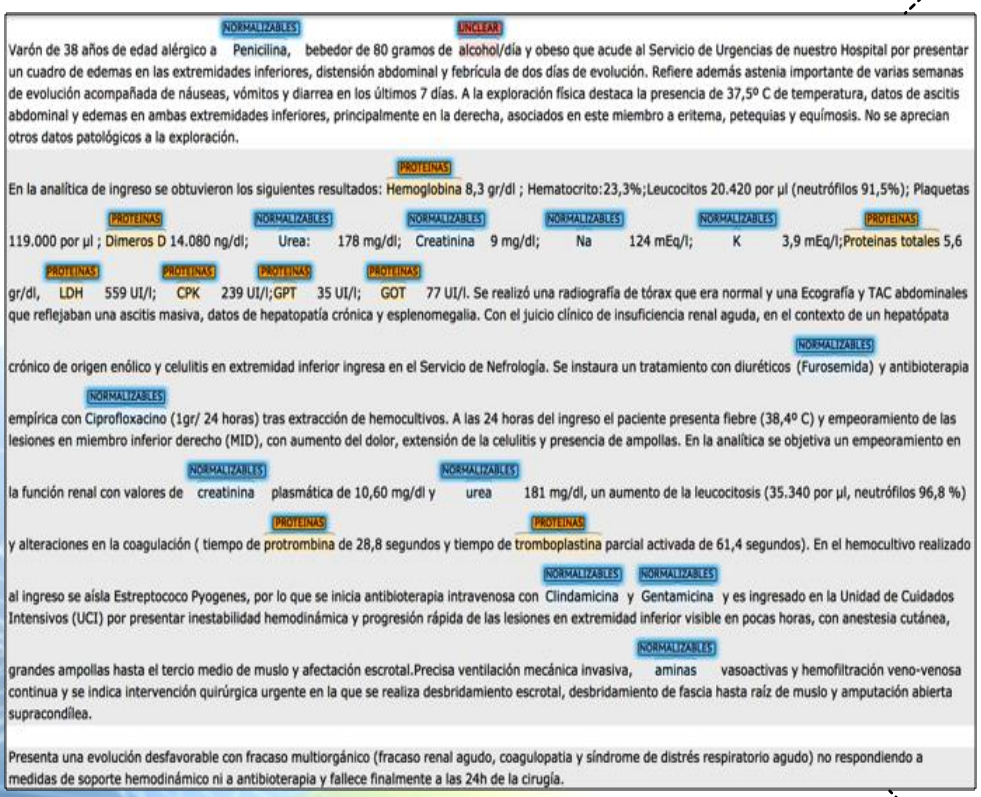
PHI entity types	PHI entity types
NOMBRE_SUJETO_ASISTENCIA	CALLE
EDAD_SUJETO_ASISTENCIA	TERRITORIO
SEXO_SUJETO_ASISTENCIA	PAIS
FAMILIARES_SUJETO_ASISTENCIA	NUMERO_TELEFONO
NOMBRE_PERSONAL_SANITARIO	NUMERO_FAX
FECHAS	CORREO_ELECTRONICO
PROFESION	ID_SUJETO_ASISTENCIA
CENTRO_SALUD	ID_CONTACTO_ASISTENCIAL
HOSPITAL	NUMERO_BENEF_PLAN_SALUD
INSTITUCION	ID_ASEGURAMIENTO
ID_TITULACION_PERSONAL_SANITARIO	URL_WEB
ID_EMPLEO_PERSONAL_SANITARIO	DIREC_PROT_INTERNET
IDENTIF_VEHICULOS_NRSERIE_PLACAS	IDENTIF_BIOMETRICOS
IDENTIF_DISPOSITIVOS_NRSERIE	OTRO_NUMERO_IDENTIF
	OTROS_SUJETO_ASISTENCIA

Mota, E., Martín, N., Moreno, A., Ferrete, E., Santamaría, J., Marimon, M., Intxaurrondo, A., Gonzalez-Agirre, A., Villegas, M., Krallinger, M.: Guías de anotación de información de salud protegida (Oct 2018).

Creación de corpus Gold Standard anotado para reconocimiento de medicamentos y compuestos

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

- **Corpus PharmaCoNER (campaña BioNLP-OST)**
  -Anotaciones de medicamentos, fármacos entidades químicas y genes/proteínas, y vacunas (incl. mapeo manual a SNOMED CT)
  -Guías de anotación y análisis de calidad (consistencia)



Varón de 38 años de edad alérgico a Penicilina, bebedor de 80 gramos de alcohol/día y obeso que acude al Servicio de Urgencias de nuestro Hospital por presentar un cuadro de edemas en las extremidades inferiores, distensión abdominal y febrícula de dos días de evolución. Refiere además astenia importante de varias semanas de evolución acompañada de náuseas, vómitos y diarrea en los últimos 7 días. A la exploración física destaca la presencia de 37,5º C de temperatura, datos de ascitis abdominal y edemas en ambas extremidades inferiores, principalmente en la derecha, asociados en este miembro a eritema, petequias y equimosis. No se aprecian otros datos patológicos a la exploración.

En la analítica de ingreso se obtuvieron los siguientes resultados: Hemoglobina 8,3 gr/dl ; Hematocrito:23,3%;Leucocitos 20.420 por µl (neutrófilos 91,5%); Plaquetas 119.000 por µl ; Dimeros D 14.080 ng/dl; Urea: 178 mg/dl; Creatinina 9 mg/dl; Na 124 mEq/l; K 3,9 mEq/l;Proteinas totales 5,6 gr/dl, LDH 559 UI/l; CPK 239 UI/l;GPT 35 UI/l; GOT 77 UI/l. Se realizó una radiografía de tórax que era normal y una Ecografía y TAC abdominales que reflejaban una ascitis masiva, datos de hepatopatía crónica y esplenomegalia. Con el juicio clínico de insuficiencia renal aguda, en el contexto de un hepatópata crónico de origen enólico y celulitis en extremidad inferior ingresa en el Servicio de Nefrología. Se instaura un tratamiento con diuréticos (Furosemida) y antibioterapia empírica con Ciprofloxacino (1gr/ 24 horas) tras extracción de hemocultivos. A las 24 horas del ingreso el paciente presenta fiebre (38,4º C) y empeoramiento de las lesiones en miembro inferior derecho (MID), con aumento del dolor, extensión de la celulitis y presencia de ampollas. En la analítica se objetiva un empeoramiento en la función renal con valores de creatinina plasmática de 10,60 mg/dl y urea 181 mg/dl, un aumento de la leucocitosis (35.340 por µl, neutrófilos 96,8 %) y alteraciones en la coagulación (tiempo de protrombina de 28,8 segundos y tiempo de tromboplastina parcial activada de 61,4 segundos). En el hemocultivo realizado al ingreso se aísla Estreptococo Pyogenes, por lo que se inicia antibioterapia intravenosa con Clindamicina y Gentamicina y es ingresado en la Unidad de Cuidados Intensivos (UCI) por presentar inestabilidad hemodinámica y progresión rápida de las lesiones en extremidad inferior visible en pocas horas, con anestesia cutánea, grandes ampollas hasta el tercio medio de muslo y afectación escrotal.Precisa ventilación mecánica invasiva, aminas vasoactivas y hemofiltración veno-venosa continua y se indica intervención quirúrgica urgente en la que se realiza desbridamiento escrotal, desbridamiento de fascia hasta raíz de muslo y amputación abierta supracondílea.

Presenta una evolución desfavorable con fracaso multiorgánico (fracaso renal agudo, coagulopatia y síndrome de distrés respiratorio agudo) no respondiendo a medidas de soporte hemodinámico ni a antibioterapia y fallece finalmente a las 24h de la cirugía.

```
T1      NORMALIZABLES 2548 2554 aminas
#1      AnnotatorNotes T1       43201005
T2      NORMALIZABLES 2223 2234 Gentamicina
#2      AnnotatorNotes T2       387321007
T3      NORMALIZABLES 2208 2220 Clindamicina
#3      AnnotatorNotes T3       372786004
T4      PROTEINAS 2034 2048     tromboplastina
#4      AnnotatorNotes T4       387124009
T5      PROTEINAS 1993 2004     protrombina
#5      AnnotatorNotes T5       7348004
T6      NORMALIZABLES 1866 1870 urea
#6      AnnotatorNotes T6       387092000
T7      NORMALIZABLES 1827 1837 creatinina
#7      AnnotatorNotes T7       15373003
T8      NORMALIZABLES 1477 1491 Ciprofloxacino
#8      AnnotatorNotes T8       372840008
T9      NORMALIZABLES 1435 1445 Furosemida
#9      AnnotatorNotes T9       387475002
T11     PROTEINAS 1029 1032     GOT
#10     AnnotatorNotes T11      26091008
T12     PROTEINAS 1016 1019     GPT
#11     AnnotatorNotes T12      56935002
T13     PROTEINAS 1003 1006     CPK
#12     AnnotatorNotes T13      75828004
T14     PROTEINAS 989 992       LDH
#13     AnnotatorNotes T14      259319003
T15     PROTEINAS 960 977       Proteinas totales
#14     AnnotatorNotes T15      395835001
T16     NORMALIZABLES 948 949   K
#15     AnnotatorNotes T16      88480006
T17     NORMALIZABLES 934 936   Na
#16     AnnotatorNotes T17      39972003
T18     NORMALIZABLES 914 924   Creatinina
#17     AnnotatorNotes T18      15373003
T19     NORMALIZABLES 897 901   Urea
#18     AnnotatorNotes T19      387092000
T20     PROTEINAS 873 882       Dimeros D
#19     AnnotatorNotes T20      25607008
T21     PROTEINAS 758 769       Hemoglobina
#20     AnnotatorNotes T21      38082009
T22     UNCLEAR 72 79   alcohol
#21     AnnotatorNotes T22      53041004
T23     NORMALIZABLES 36 46     Penicilina
#22     AnnotatorNotes T23      323389000
```

ciprofloxacino (sustancia)
SCTID: 372840008
372840008 | ciprofloxacino (sustancia) |
en Ciprofloxacin (substance)
en Ciprofloxacin
es ciprofloxacina
es ciprofloxacino (sustancia)
es ciprofloxacino

cima
CENTRO DE INVESTIGACIÓN MÉDICA APLICADA
UNIVERSIDAD DE NAVARRA

http://temu.bsc.es/pharmaconer/

Creación de corpus Gold Standard de entidades nombradas de relevancia clinica

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Corpus de entidades de interés clínico en español

-Anotaciones de conceptos y entidades nombradas médicas asociadas a los siguientes tipos: enfermedad/síntoma, procedimiento, fármaco y entidad observable. Mapeo manual de entidades a SNOMED CT.
-Guías de anotación y análisis de calidad (consistencia)
-Fase de revisión final y definición de shared task (campaña de evaluación)

Paciente de 70 años de edad, minero jubilado, sin alergias medicamentosas conocidas, que presenta como antecedentes personales: accidente laboral antiguo con fracturas vertebrales y costales; intervenido de enfermedad de Dupuytren en mano derecha y by-pass iliofemoral izquierdo; Diabetes Mellitus tipo II, hipercolesterolemia e hiperuricemia; enolismo activo, fumador de 20 cigarrillos / día.

Es derivado desde Atención Primaria por presentar hematuria macroscópica postmiccional en una ocasión y microhematuria persistente posteriormente, con micciones normales.

En la exploración física presenta un buen estado general, con abdomen y genitales normales; tacto rectal compatible con adenoma de próstata grado I/IV.

En la analítica de orina destaca la existencia de 4 hematíes/ campo y 0-5 leucocitos/campo; resto de sedimento normal.

Hemograma normal; en la bioquímica destaca una glucemia de 169 mg/dl y triglicéridos de 456 mg/dl; función hepática y renal normal. PSA de 1.16 ng/ml.

Las citologías de orina son repetidamente sospechosas de malignidad.

En la placa simple de abdomen se valoran cambios degenerativos en columna lumbar y calcificaciones vasculares en ambos hipocondrios y en pelvis.

La ecografía urológica pone de manifiesto la existencia de quistes corticales simples en riñón derecho, vejiga sin alteraciones con buena capacidad y próstata con un peso c 30 g.

En la UIV se observa normofuncionalismo renal bilateral, calcificaciones sobre silueta renal derecha y uréteres arrosariados con imágenes de adición en el tercio superior c ambos uréteres, en relación a pseudodiverticulosis ureteral. El cistograma demuestra una vejiga con buena capacidad, pero paredes trabeculadas en relación a vejiga de esfuerzo. La TC abdominal es normal.

La cistoscopia descubre la existencia de pequeñas tumoraciones vesicales, realizándose resección transuretral con el resultado anatomopatológico de carcinoma urotelial superficial de vejiga.

- SINTOMA
- ENFERMEDAD
- PROCEDIMIENTO_DIAGNOSTICO
- PROCEDIMIENTO_TERAPEUTICO
- FARMACO

VeraTech FOR HEALTH

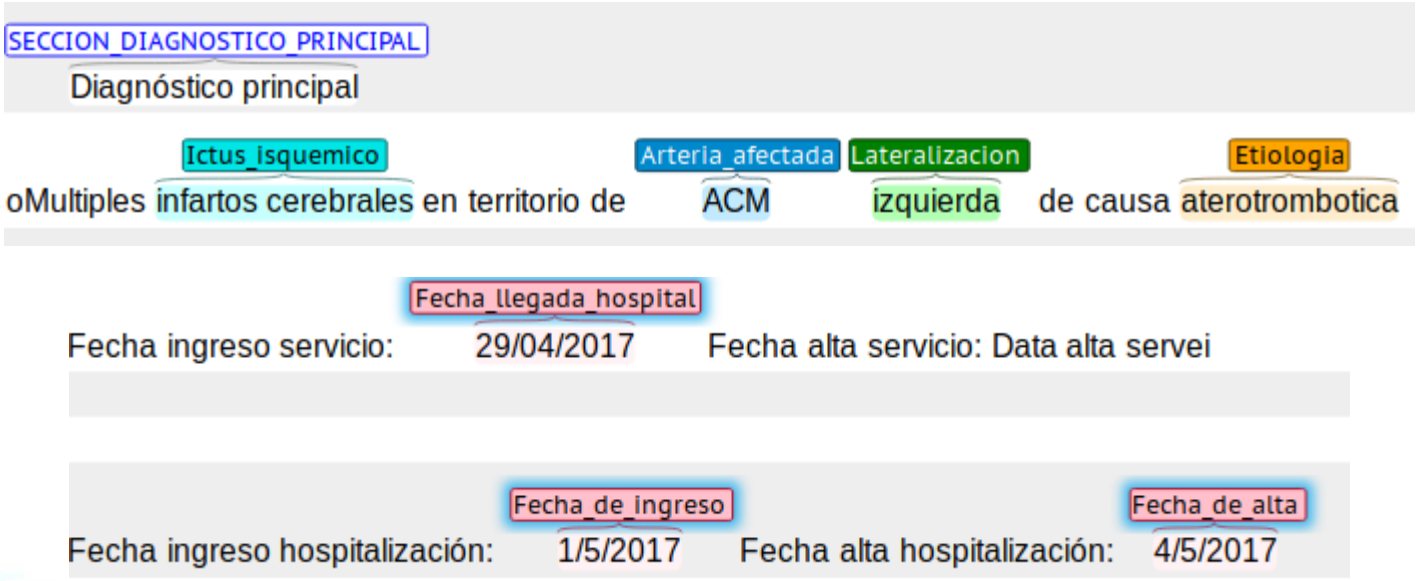
Hospital Universitario 12 de Octubre

Creación de corpus Gold Standard anotado con expresiones temporales y secciones clinicas

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Corpus anotados por el PlanTL en colaboración con consorcio IctusNet

- Anotación de expresiones y eventos temporales, indicadores relacionados con el ictus y secciones de los informes de alta (mapeo a arquetipos)
- Pre-anotación automática, guías de anotación y corrección manual, análisis de calidad.



Interreg Sudoe
EUROPEAN UNION
ICTUSnet
European Regional Development Fund

Guias o esquemas de anotacion: clave para interpretar, reproducir, expandir, adaptar

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

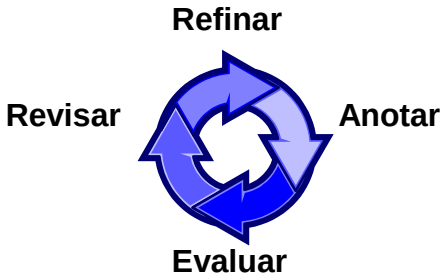
Guías o esquemas de anotación (reglas y criterios de etiquetado manual)

- Proporcionan detalles básicos de anotación a seguir durante el proceso de construcción del corpus.

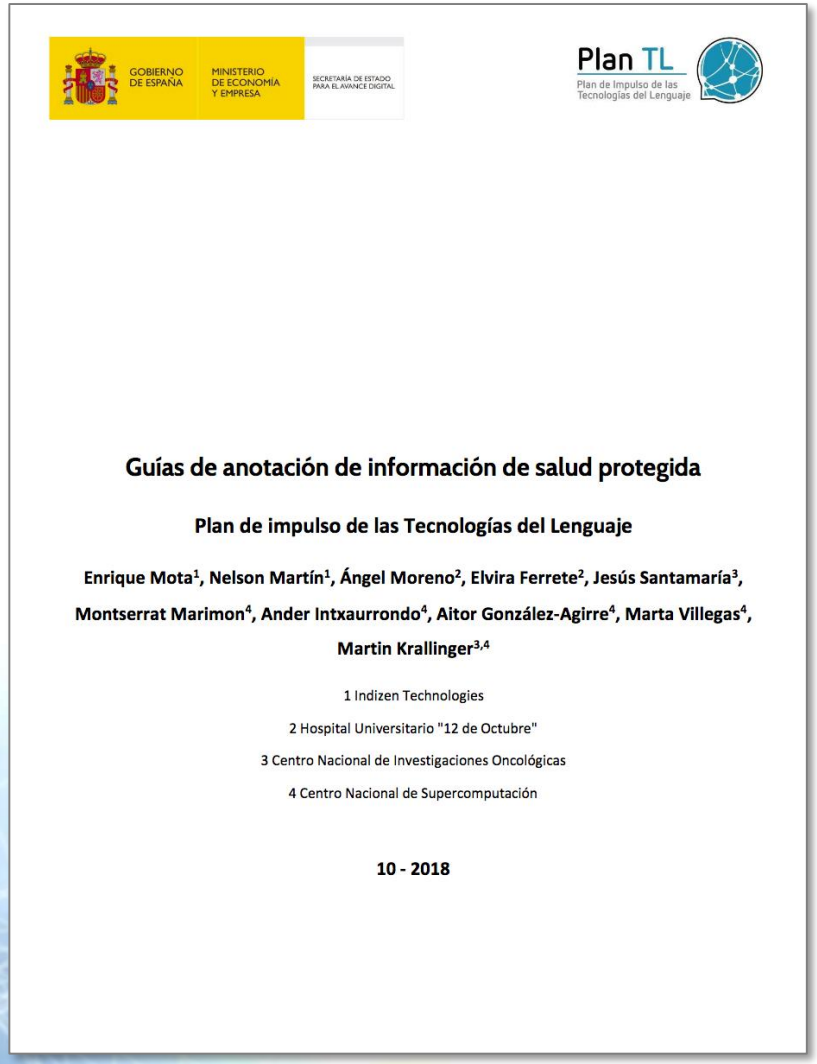
- Ayudan a los anotadores a producir datos consistentes y a los usuarios finales a interpretar las anotaciones correctamente.

- Se refinan tras sucesivos ciclos iterativos de anotación de documentos de muestra, basadas en las sugerencias directas realizadas por los anotadores, así como por la observación de inconsistencias detectadas al comparar los resultados de diferentes anotadores.

- Reglas de anotación:
  - Reglas-positivas: especifican qué se debe anotar
  - Reglas-negativa: especifican qué no se debe anotar
  - Reglas-multipalabra: reglas que especifican qué grupo de palabras debe anotarse bajo una única etiqueta.
  - ...



Guias o esquemas de anotacion: clave para interpretar, reproducir, expandir, adaptar



Guías de anotación de información de salud protegida

Plan de impulso de las Tecnologías del Lenguaje

Enrique Mota[1], Nelson Martín[1], Ángel Moreno[2], Elvira Ferrete[2], Jesús Santamaría[3], Montserrat Marimon[4], Ander Intxaurrondo[4], Aitor González-Agirre[4], Marta Villegas[4], Martin Krallinger[3,4]

1 Indizen Technologies

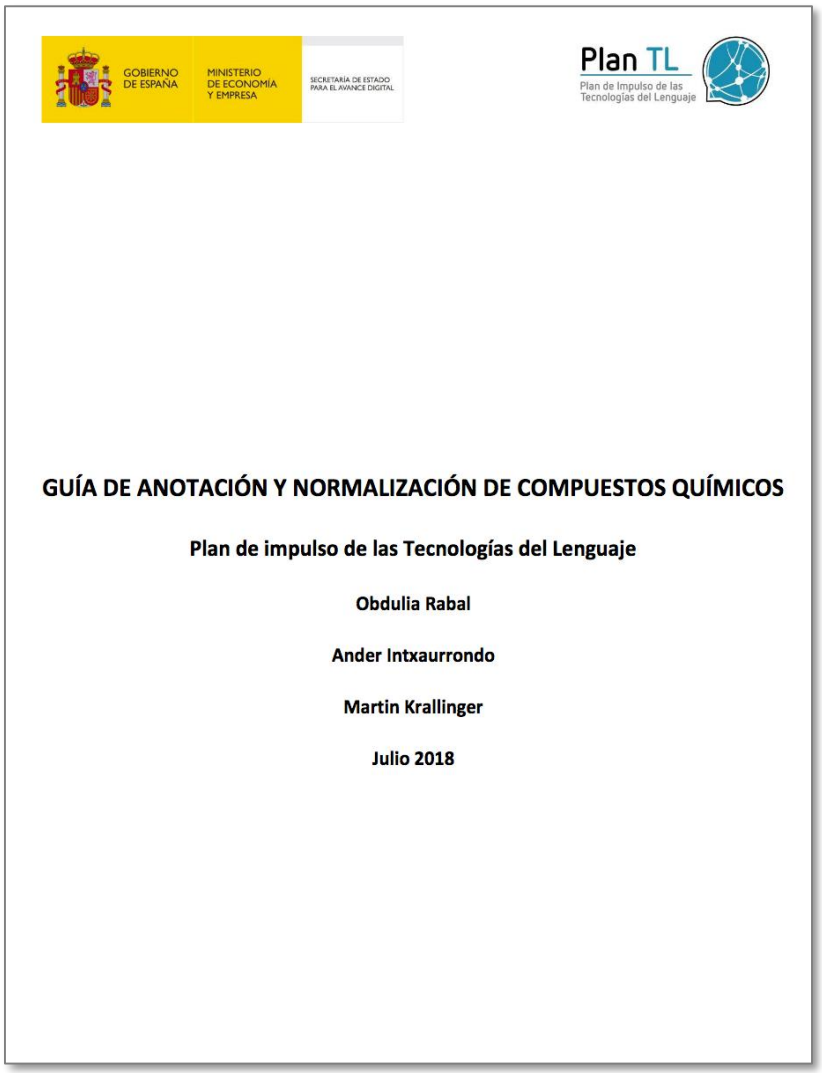
2 Hospital Universitario "12 de Octubre"

3 Centro Nacional de Investigaciones Oncológicas

4 Centro Nacional de Supercomputación

10 - 2018

http://temu.bsc.es/meddocan



GUÍA DE ANOTACIÓN Y NORMALIZACIÓN DE COMPUESTOS QUÍMICOS

Plan de impulso de las Tecnologías del Lenguaje

Obdulia Rabal

Ander Intxaurrondo

Martin Krallinger

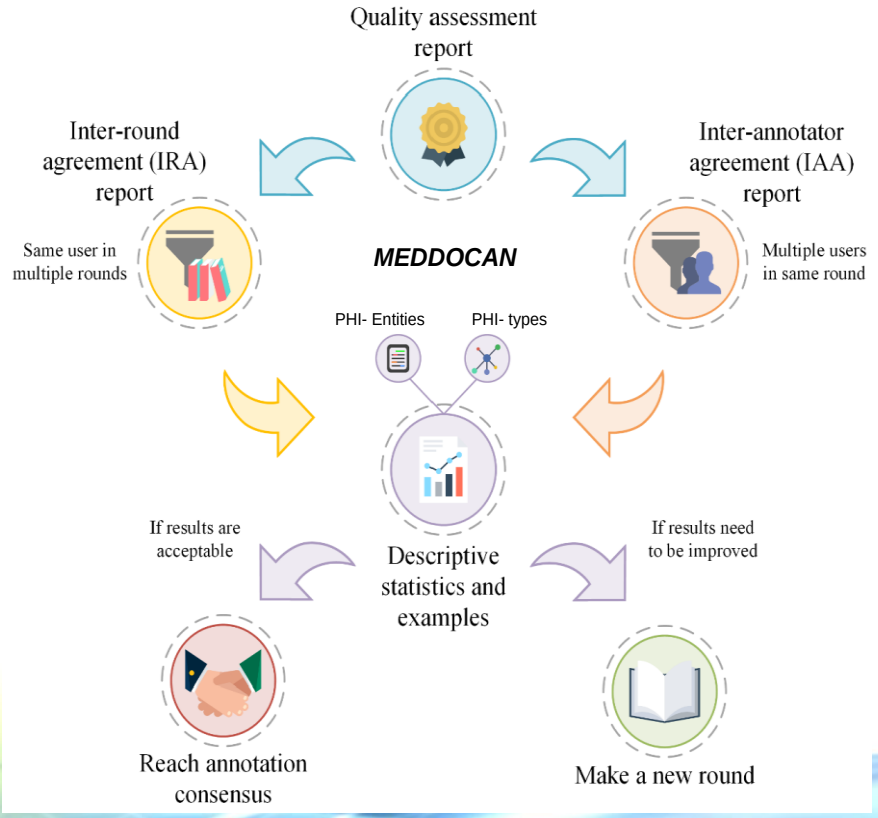
Julio 2018

http://temu.bsc.es/pharmaconer

Creación de corpus anotados: control de calidad (inter-anotator agreement)

Control de la calidad

- Incluso con unas buenas guías de anotación, las anotaciones humanas no serán perfectas.
- Análisis de coherencia de anotación mediante el acuerdo entre anotadores (IAA, del inglés Inter-Annotator Agreement).



Creación de corpus silver standard anotados automaticamente por multiples sistemas

Campañas de evaluación para creación de corpus silver standard

- Creación de Silver Standard: anotación automática por sistemas participantes en campañas de evaluación: MEDDOCAN Silver Standard para **2,751 nuevos documentos**

- Anotación automática a través de tarea MEDDOCAN para:
  - Identificación y clasificación de entidades (63 sistemas)
  - Detección de texto sensible (61 sistemas)
  - Resultados:
    - **0,96961 y 0,97491 (F1)**

Creación de corpus silver standard anotados automaticamente por multiples sistemas

Campañas de evaluación para creación de corpus silver standard

- Creación de Silver Standard: anotación automática por sistemas participantes en campañas de evaluación: PharmaCoNER Silver Standard para **2,751 nuevos documentos**

- Anotación automática a través de tarea PharmaCoNER para:
  - Identificación y clasificación de medicamentos, compuestos y genes (76 sistemas)
  - Indización de documentos con SNOMED CT (19 sistemas)
- Resultados:
  - **0,91052 y 0,91593 (F1)**

Creación de corpus Social Media para Salud (Twitter)

Campañas de evaluación: Social Media Mining for Health Applications (#SMM4H) Shared Task 2020

- Corpus para el reconocimiento de efectos adversos a medicación en twitter
- Corpus de entrenamiento (training) data: ~4,000 tweets
- Corpus de evaluación (test): ~1,000 tweets

https://healthlanguageprocessing.org/smm4h-sharedtask-2020/

Gracias

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- Martin Krallinger
- Marta Villegas
- Siamak Barzegar
- Antonio Miranda
- Alejandro Asensio
- Aitor Gonzalez
- Montse Marimon
- Felipe Soares

- Alfonso Valencia (BSC Life)

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

David Perez (SEAD)

- *AQuAS* (Miguel Gallofre López)
- *AEMPS-BIFAP* (Julio Bonis Sanz)
- AEMPS-FTM (JM Simarro)
- *FID-Salud/MSSSI* (Elena García)
- *FISEVI/Hosp. Virgen del Rocio* (Carlos Parra)
- *Hospital 12 de Octubre* (Pablo Serrano)
- *IBECS/Carlos III* (Elena Primo)
- *Informática Médica Hosp. Clínic* (Raimundo Lozano)
- *MSSSI* (Maribel García Fajardo)
- *RANM* (Cristina V. González)

cima
CENTER FOR APPLIED MEDICAL RESEARCH
UNIVERSITY OF NAVARRA

- Obdulia Rabal
- Julen Oyarzabal

BIO CREATIVE

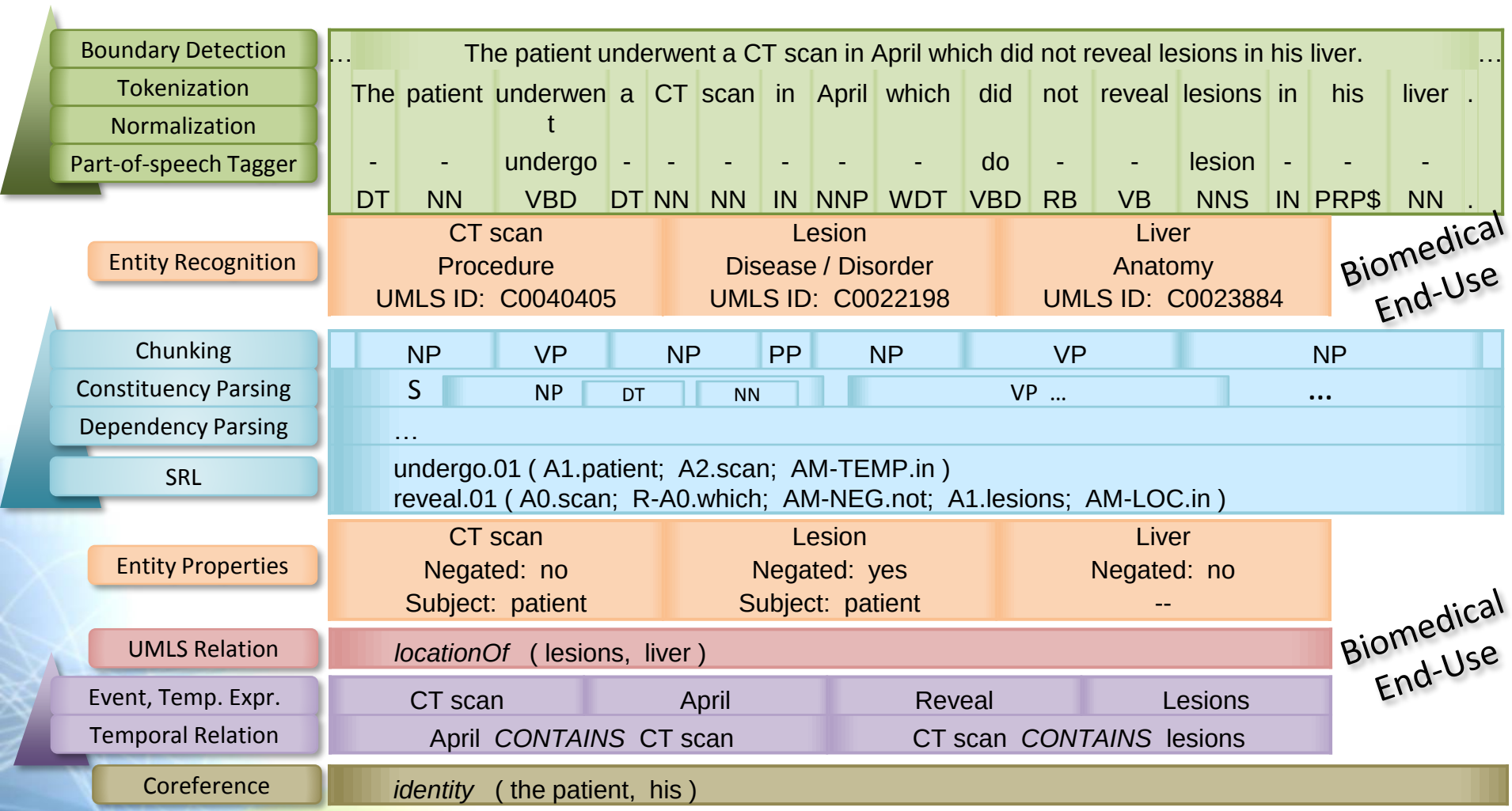
- BioCreative organizers
- Cecilia Arighi/Cathy Wu (Uni. Delaware)
- Lynette Hirschman (MITRE)

Universidade Vigo

- Analia Lourenço
- Martin Perez Perez
- Gael Perez Rodriguez
- Florentino Fernández Riverola

Anotaciones relevantes para PLN clínico: cTakes

The patient underwent a CT scan in April which did not reveal lesions in his liver.



1. Proceso de anotación y text mining

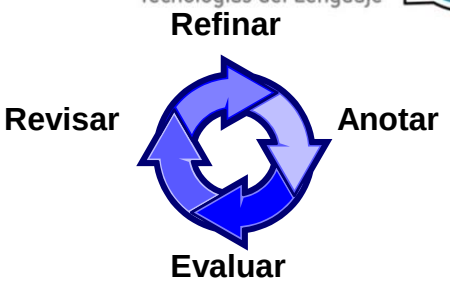
Documentos

Papers científicos
Patentes
Informes clínicos

Definir qué se quiere anotar y con qué propósito

Guías de anotación

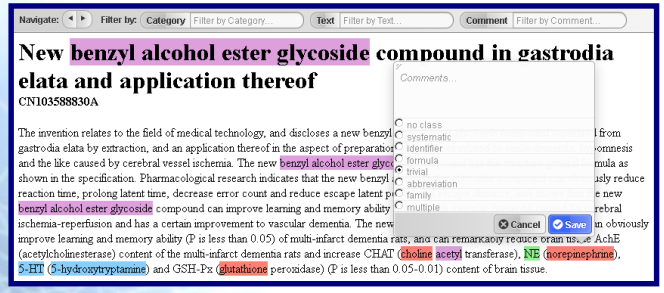
Especifican qué y qué no anotar
Definen reglas consensuadas de cómo anotar la información



1º) Realizar una ronda de anotación (sample test) las personas que definen las guías

2º) Una vez definidas guías + sample test → pasar a anotadores

Anotación manual de textos



Training set

Development set

Generación de la herramienta de text mining

Test set

Evaluación de la herramienta de text mining

Patent	Part	Start	End	CEM	CLASS
CN103588830A	T	4	24	benzyl alcohol ester	SYSTEMATIC
CN103588830A	A	1189	1190	NE	ABBREVIATION
CN103588830A	A	1193	1206	norepinephrine	TRIVIAL

Generación de las guías de anotación

1. Revisar guías de anotación previas publicados para las entidades que se quiere anotar (si las hay)

2. Revisar 1-10 documentos para ver ejemplos reales de cómo quedarían anotados
   (da idea de la casuística inicial que puede ir surgiendo)

3. Estructurar las guías como reglas:

   - Generales: qué debe hacer el anotador en caso de duda, recursos que puede consultar (DBs, webs…)
   - Positivas: reglas que definan las entidades que se tienen que anotar
   - Negativas: reglas que definan las entidades que NO se deben anotar
   - Clases: reglas de clasificación de entidades en diferentes subgrupos (si aplica y es útil)
   - Ortografía: qué hacer con palabras en otros idiomas, errores tipográficos, signos de puntuación
   - Palabras múltiples:
     - qué hacer con adjetivos que acompañan a entidades a anotar? “fiebre aguda”
     - qué hacer con adjetivos que son negativos “no hiperglicemia”
     - enumeraciones, listas de entidades: anotar de manera separada o conjunta?

       glucosa, colesterol, ..
       glucosa-colesterol *???*

4. Acompañar las reglas en las guías con ejemplos diversos → incorporar ejemplos conforme se vayan refinando

5. Conviene que las reglas sean consensuadas por > 1 persona y también revisadas por alguién “externo” a su escritura para ver si se entienden / son coherentes

6. No muy largas (20 – 30 páginas)…

Training anotadores, actualización guías, revisión anotaciones

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1. Aplicar las guías estrictamente para anotar una serie de documentos (sample-set): en este momento irán saliendo las primeras dudas y se puede refinan sin tener que avisar a anotadores: incluir más ejemplos, decidir entre varias personas el consenso (los términos frontera son los más problemáticos)

2. Una vez anotados y con una versión cuasi-definitiva de las guías: pasárselas a anotadores

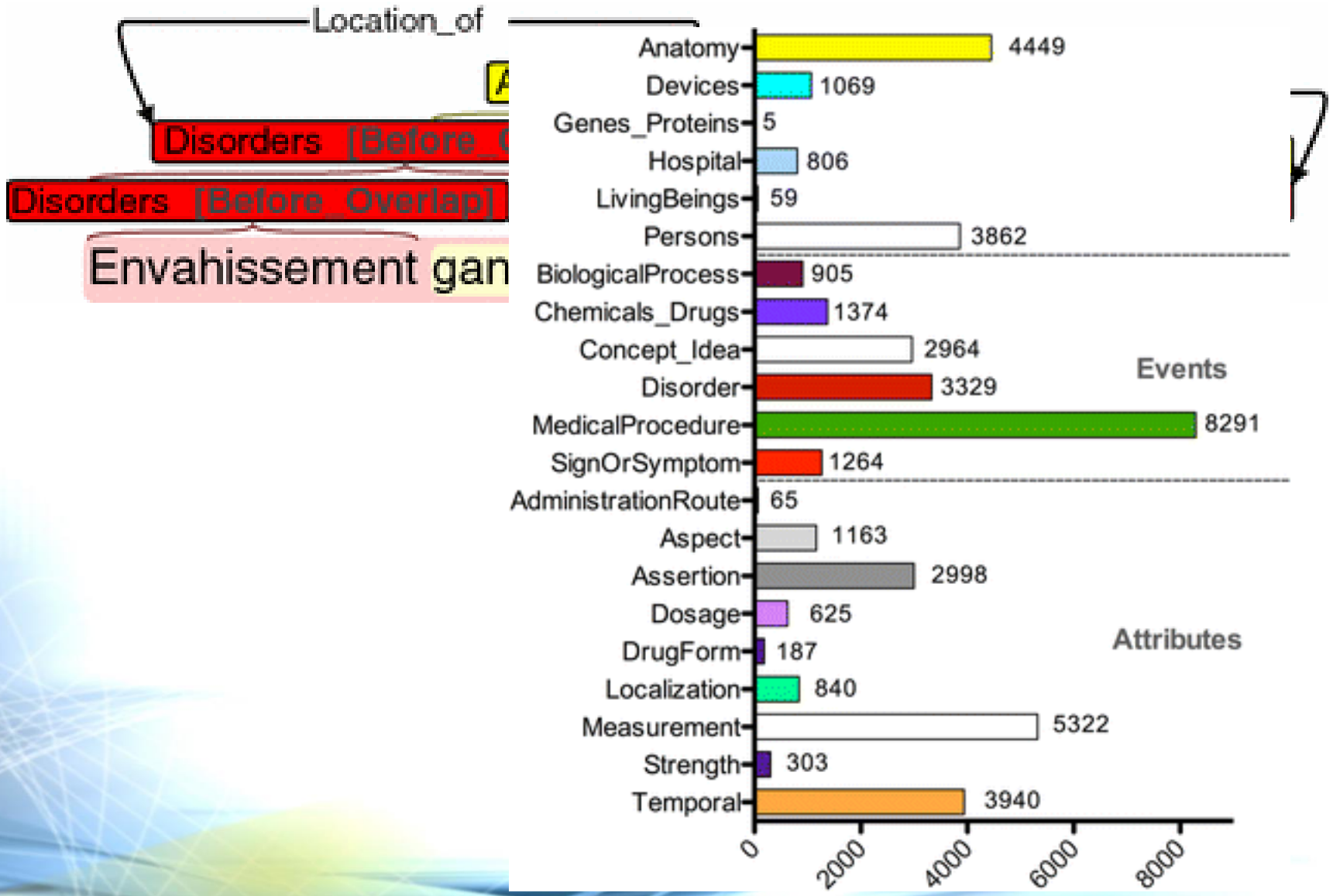
3. Anotadores: personas con experiencia en el campo de la entidad que se va a anotar (enfermedades, síntomas)

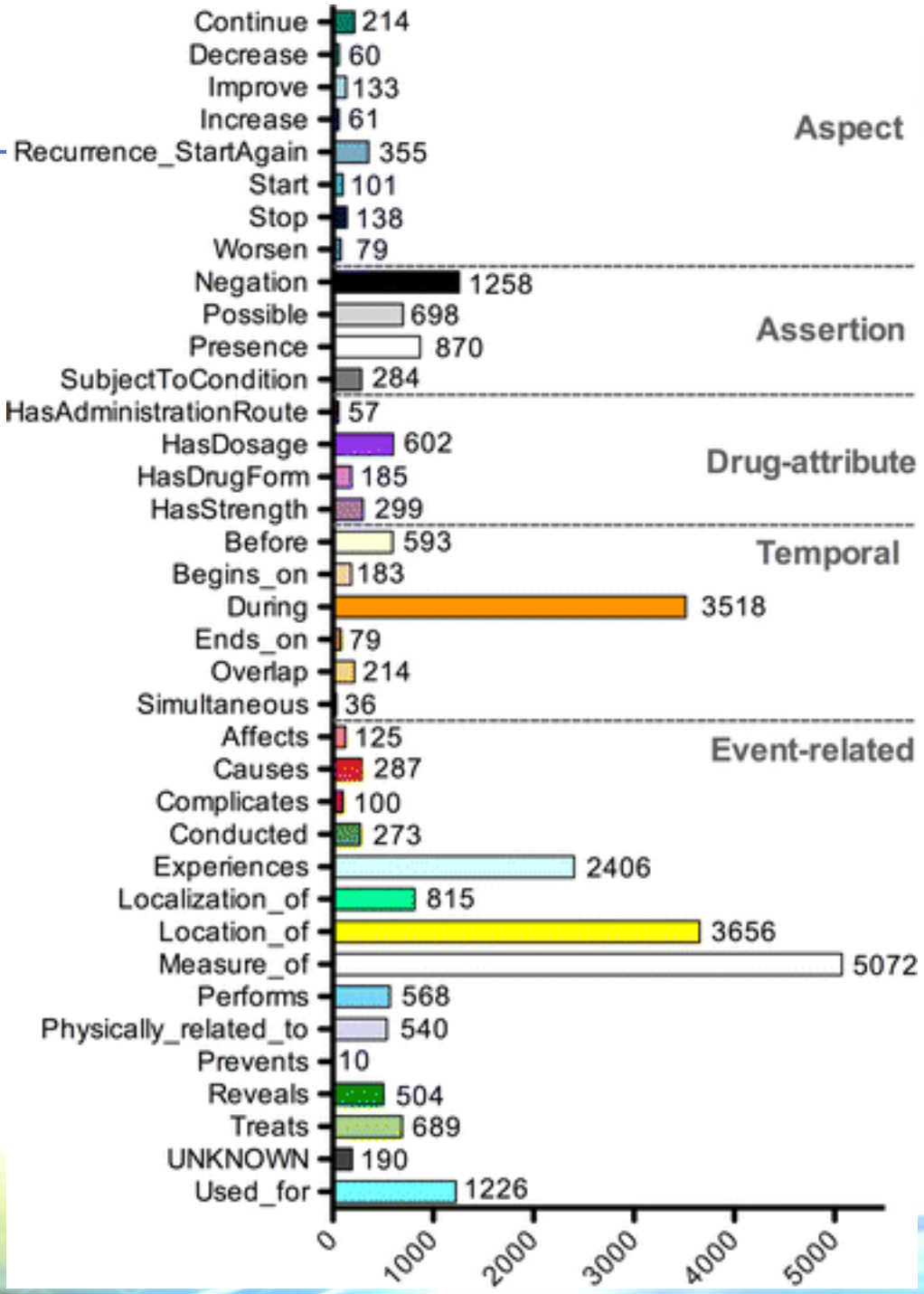
4. Training a anotadores:

   - Versión final de las guías de anotación
   - Demo o manual de cómo funciona la herramienta de anotación
   - El sample set sirve como comparativa para ver la relación anotadores – guías (exhaustividad, comprensión)
   - Persona que genere las guías que sirva de contacto en caso de duda
   - Si durante el proceso de anotación se detecta un problema con las guías → actualizar guías

5. Revisión de las anotaciones, comparación entre anotadores

Sample annotation from the MERLOT corpus




Continue 214
Decrease 60
Improve 133
Increase 61
Recurrence_StartAgain 355
Start 101
Stop 138
Worsen 79
Aspect
Negation 1258
Possible 698
Presence 870
SubjectToCondition 284
Assertion
HasAdministrationRoute 57
HasDosage 602
HasDrugForm 185
HasStrength 299
Drug-attribute
Before 593
Begins_on 183
During 3518
Ends_on 79
Overlap 214
Simultaneous 36
Temporal
Affects 125
Causes 287
Complicates 100
Conducted 273
Experiences 2406
Localization_of 815
Location_of 3656
Measure_of 5072
Performs 568
Physically_related_to 540
Prevents 10
Reveals 504
Treats 689
UNKNOWN 190
Used_for 1226
Event-related
0
1000
2000
3000
4000
5000

Corpus clínicos y biomédicos en español

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Costumero R, García-Pedrero A, Gonzalo-Martín C, Menasalvas E, Millan S. Text analysis and information extraction from Spanish written documents In: Slezak D, Tan A. -H, Peters J, Schwabe L, editors. Brain Informatics and Health. Lecture Notes in Computer Science. Springer: 2014. p. 188–197.

Segura-Bedmar I, de la Peña González S, Martínez P. Extracting drug indications and adverse drug reactions from Spanish health social media. In: Proceedings of BioNLP 2014. Baltimore, Maryland: Association for Computational Linguistics: 2014. p. 98–106. http://www.aclweb.org/anthology/W/W14/W14-3415.

Costumero R, Lopez F, Gonzalo-Martín C, Millan M, Menasalvas E. An Approach to Detect Negation on Medical Documents in Spanish. In: International Conference on Brain Informatics and Health. Springer: 2014. p. 366–375.

Figueroa R, Soto D, Pino E. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. In: Conf Proc IEEE Eng Med Biol Soc: 2014. p. 2710–3.

Corpus clínicos y biomédicos en español

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Kors J, Clematide S, Akhondi S, van Mulligen E, Rebholz-Schuhmann D. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. J Am Med Inform Assoc. 2015; 22(5):948–56.

Oronoz M, Gojenola K, Pérez A, de Ilarraza A, A AC. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. J Biomed Inform. 2015; 56:318–32.

Castano J, Gambarte ML, Park HJ, Avila Williams MdP, Perez D, Campos F, Luna D, Benitez S, Berinsky H, Zanetti S. A machine learning approach to clinical terms normalization. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics: 2016. p. 1–11.

Cotik V, Stricker V, Vivaldi J, Rodriguez H. Syntactic methods for negation detection in radiology reports in Spanish. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics: 2016. p. 156–165.

Corpus clínicos y biomédicos en español

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Neves M, Yepes AJ, Névéol A. The scielo corpus: a parallel corpus of scientific publications for biomedicine In: Chair NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris, France: European Language Resources Association (ELRA): 2016.

Liu W, Cai S. Translating electronic health record notes from English to Spanish: A preliminary study. In: Proceedings of BioNLP 15. Beijing, China: Association for Computational Linguistics: 2015. p. 134–140. http://www.aclweb.org/anthology/W15-3816.

Creación de corpus Gold Standard para reconocimiento de entidades de carácter personal

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Type	Train	Dev	Test	Total
TERRITORIO	1875	987	956	3818
FECHAS	1231	724	611	2566
EDAD_SUJETO_ASISTENCIA	1035	521	518	2074
NOMBRE_SUJETO_ASISTENCIA	1009	503	502	2014
NOMBRE_PERSONAL_SANITARIO	1000	497	501	1998
SEXO_SUJETO_ASISTENCIA	925	455	461	1841
CALLE	862	434	413	1709
PAIS	713	347	363	1423
ID_SUJETO_ASISTENCIA	567	292	283	1142
CORREO_ELECTRONICO	469	241	249	959
ID_TITULACION_PERSONAL_SANITARIO	471	226	234	931
ID_ASEGURAMIENTO	391	194	198	783
HOSPITAL	255	140	130	525
FAMILIARES_SUJETO_ASISTENCIA	243	92	81	416
INSTITUCION	98	72	67	237
ID_CONTACTO_ASISTENCIAL	77	32	39	148
NUMERO_TELEFONO	58	25	26	109
PROFESION	24	4	9	37
NUMERO_FAX	15	6	7	28
OTROS_SUJETO_ASISTENCIA	9	6	7	22
CENTRO_SALUD	6	2	6	14
ID_EMPLEO_PERSONAL_SANITARIO	0	1	0	1
IDENTIF VEHICULOS NRSERIE PLACAS	0	0	0	0
IDENTIF DISPOSITIVOS NRSERIE	0	0	0	0
NUMERO BENEF PLAN SALUD	0	0	0	0
URL WEB	0	0	0	0
DIREC PROT INTERNET	0	0	0	0
IDENTF BIOMETRICOS	0	0	0	0
OTRO NUMERO IDENTIF	0	0	0	0

MEDDOCAN track proceedings (IberLEF)

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Track 9: Medical Document Anonymization Task (MEDDOCAN)

Formatos y esquemas de anotación en PLN

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Modelo	Dominio	Formato de serializacion	API	Tipo
BioC	Biomedical	XML	Reference APIs in multiple languages	Stand-off
BioNLP shared task TSV	Biomedical	TSV	No	Stand-off
BRAT format	Generic	TSV	No	Stand-off
Pubtator	Biomedical	TSV	No	Stand-off
TEI	Generic	XML	Via XSLT	Stand-off
NIF	Generic	RDF	No	Stand-off
LIF	Generic	RDF	Reference API in Java	Stand-off
IOB	Generic	TSV	Third-party APIs in several languages	In-line
Open Annotation	Generic	RDF	No	Stand-off
CAS (UIMA)	Generic	XML (XMI)	Reference APIs in Java and C ++	Stand-off and in-line
GATE annotation format	Generic	Several	Reference API in Java	Stand-off and in-line
LAF/GrAF	Generic	XML	No	Stand-off
PubAnnotation	Generic	JSON	REST API to annotation store	Stand-off

Principales formatos y esquemas de anotación en PLN con especial énfasis en los usados en el dominio biomédico

Principales esquemas de metadatos para contenidos

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Nombre	Dominio	Uso	URL
Dublin Core (DC)/DC Metadata Initiative (DCMI)	Genérico	Estándar ampliamente aceptado	http://dublincore.org/
Journal Article Tag Suite (JATS)	Publicaciones Científicas	Revistas Open Access	https://jats.nlm.nih.gov/
DataCite	Datos de investigación y publicaciones	Citations	https://www.datacite.org/
CrossRef	Datos de investigación y publicaciones	Referencias	http://www.crossref.org/
BibJSON	Información bibliográfica	Metadatos de Referencias	http://okfnlabs.org/bibjson/
CERIF	Información de Investigación europea	Investigación	http://www.eurocris.org/cerif/main-features-cerif
CKANg	Genérico/Data management	Portales	http://ckan.org/