GUÍA DE ANOTACIÓN DE TEXTOS MÉDICOS EN ESPAÑOL:

ANOTACIÓN CATEGORIAL

Plan de impulso de las Tecnologías del Lenguaje

Nuria Aldama García[1]

Carmen Torrijos Caruda[1]

Montserrat Marimon[2]

Martin Krallinger[2,3]

[1]Instituto de ingeniería del conocimiento
[2]Centro Nacional de Supercomputación
[3]Centro nacional de Investigaciones Oncológicas

Julio 2018

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

ÍNDICE DE TABLAS

RESUMEN

Este documento presenta la herramienta utilizada para la anotación categorial de textos médicos en español, así como las convenciones que deben ser seguidas durante el proceso de anotación manual del corpus.

1 INTRODUCCIÓN

El Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) tiene como objetivo fomentar el desarrollo del Procesamiento del Lenguaje Natural (PLN) y la Traducción Automática (TA) en lengua española y lenguas cooficiales. Para ello, el Plan TL define medidas que:

- Aumenten el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas cooficiales.
- Impulsen la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria.
- Incorporen a la Administración como impulsor del sector de PLN.

Uno de los objetivos del proyecto es poner a disposición de la comunidad científica y la industria un corpus biomédico exhaustivo y con licencia abierta que permita ejecutar tareas de PLN sobre *big data* y replicar los experimentos. Este documento presenta la herramienta utilizada para la anotación categorial de textos médicos en español, así como las convenciones que deben ser seguidas durante el proceso de anotación manual del corpus. Consultar el documento *Metodología de anotación de textos biomédicos en español* para conocer los detalles relativos a los perfiles del autor de la guía y de los anotadores del corpus.

2 FREELING3.1

FreeLing [9] es una herramienta de análisis y etiquetado lingüístico que permite identificar el lenguaje al que pertenece una expresión lingüística, dividirla en oraciones, lematizarla y etiquetarla morfosintácticamente. Es una aplicación de código abierto para el procesamiento automático del lenguaje natural que proporciona una amplia gama de servicios de análisis lingüístico para una gran variedad de idiomas. Esta librería es personalizable y ampliable, y está fuertemente orientada al desarrollo de aplicaciones del mundo real en términos de velocidad y robustez. Además, permite al usuario analizar archivos de texto desde la línea de comandos. Por estos motivos, FreeLing es la herramienta que hemos elegido para la anotación de textos médicos en español, creando una

versión mejorada y adaptada al dominio médico a través de la modificación y el enriquecimiento de sus recursos de base.

2.1 RECURSOS POR DEFECTO (FREELING BASELINE)

Para realizar la asignación del Part of Speech (POS), FreeLing3.1 contiene una serie de recursos que se detallan a continuación y que establecen el *baseline*:

2.2 FICHERO DE SINGLE WORDS BÁSICO DE FREELING3.1

Incluye un total de 669.291 entradas de palabras básicas. En la siguiente tabla se detalla la distribución de las categorías morfológicas, el POS con el que se identifican y su número de apariciones:

Categoría	***POS***	***Apariciones***
Adjetivos	*AQ / AO*	*61936 / 993*
Conjunciones	*CC / CS*	*18 / 19*
Determinantes	*DA / DD / DE / DI / DP / DT*	*5 / 24 / 7 / 52 / 14 / 7*
Puntuación	*Fs*	*3*
Interjección	*I*	*201*
Sustantivo	*NC*	*107655*
Pronombre	*P0 / PD / PI / PP / PR / PT / PX*	*8 / 38 / 13 / 61 / 31 / 17 / 13 / 25*
Adverbio	*RG / RN*	*184 / 1*
Preposición	*SP*	*30*
Verbo	*VM / VS / VA*	*497756 / 62 / 118*
TOTAL		*669.291*

Tabla 1: Distribución de las palabras básicas por categorías morfológicas, el POS con el que se identifican y su número de apariciones.

Estas palabras están normalizadas en los ficheros MM* de la carpeta dictionary/. Estos ficheros se estructuran en tres columnas con la siguiente cabecera:

Forma	Lema	POS

2.3 FICHERO DE MULTIWORDS BÁSICO DE FREELING3.1

Incluye un total de 3.124 entradas de palabras complejas. En la siguiente tabla se detalla la distribución de las categorías morfológicas:

Categoría	***POS***	***Apariciones***
Adjetivo	*AQ*	*18*
Conjunción	*CC / CS*	*7 / 50*
Determinante	*DD*	*1*
Puntuación	*Fs*	*1*
Interjección	*I*	*6*
Sustantivo	*NC*	*1677*
Adverbio	*RG*	*1155*
Preposición	*SP*	*213*
Verbo	*VM / VS*	*35 / 1*
TOTAL		*3164*

Tabla 2: Distribución de las palabras complejas por categorías morfológicas, el POS con el que se identifican y su número de apariciones.

Estas palabras están normalizadas en los ficheros locucions.dat y locucions_extended.dat. Estos ficheros se estructuran en tres columnas con la siguiente cabecera:

Forma	Lema	POS

2.4 ETIQUETARIO MORFOLÓGICO DE FREELING3.1

Para el etiquetado morfológico (POS) FreeLing utiliza las etiquetas EAGLES – PAROLE (http://www.lsi.upc.es/~nlp/tools/parole-sp.html), propuestas por el grupo EAGLES (Expert Advisory Group on Language Engineering Standards) para la anotación morfosintáctica de lexicones y corpus. Está previsto que recoja los accidentes gramaticales existentes en todas las lenguas europeas. Para cada categoría gramatical se presentan los atributos, valores y códigos que puede tomar. Dependiendo de la lengua, hay atributos que pueden no especificarse, en cuyo caso se marcan con un 0. A continuación se señala un ejemplo por cada categoría gramatical:

Forma	***POS***	***Desarrollo del POS***
Reunión	*NCFS000*	*Nombre común femenino singular*
Simpático	*AQ0MS0*	*Adjetivo calificativo masculino singular*
Para	*SPS00*	*Sintagma preposicional*
Despacio	*RG*	*Adverbio*
Comemos	*VMIP1P0*	*Verbo en presente de indicativo y primera persona del plural*
La	*TDFS0*	*Determinante artículo femenino singular*
Aquellos	*DD3MP00*	*Determinante demostrativo tercera persona masculino plural*
Vuestras	*DP3FP05*	*Determinante posesivo femenino plural segunda persona*
Ni	*CC*	*Conjunción coordinada*
Tercer	*MOMS00*	*Numeral ordinal masculino singular*
Ah	*I*	*Interjección*
.	*Fp*	*Puntuación*
etc	*Y*	*Abreviatura*

Tabla 3: Ejemplos del etiquetario morfológico de FreeLing3.1.

Este etiquetario se utilizará tanto en la anotación manual como en la anotación automática y en la visualización de BRAT.

3 REGLAS DE ANOTACIÓN MANUAL

Estas reglas proporcionan los detalles básicos de la anotación y las convenciones que deben ser seguidas durante el proceso de anotación manual del corpus. Las reglas se dividen en:

- Reglas generales: Reglas básicas que aplican a todos los procedimientos de etiquetado morfológico.
- Reglas positivas: Reglas que aplican a casos específicos con una determinada etiqueta morfológica. Se acompañan de ejemplos.
- Reglas negativas: Reglas que aplican a casos específicos que no deben etiquetarse morfológicamente. Se acompañan de ejemplos.
- Reglas ortográficas: Reglas que aplican a errores de ortotipografía. Se acompañan de ejemplos.

3.1 REGLAS GENERALES (REGLAS-G)

- **G1. Género**

Cuando el género gramatical de palabras relativas a términos médicos no es identificable se anota con el valor 0.

Ejemplo:

Forma	***Lema***	***POS***
cetuximab	*cetuximab*	*NC00000*

- **G2. Número**

Cuando el número gramatical de palabras relativas a términos médicos no es identificable se anota con el valor 0.

Ejemplo:

Forma	***Lema***	***POS***
bolus	*bolus*	*NC00000*

- **G3. Desarrollo de lemas de siglas y abreviaturas**

El lema de las siglas y abreviaturas no se desarrolla. Si el género y el número no son identificables se anotarán con el valor 0.

Ejemplo:

Forma	***Lema***	***POS***
LSI	*lsi*	*NC00000*
Sra.	*sra*	*NCFS000*

- **G4. Unidades de medida**

El lema de las unidades de medida se desarrolla y se etiqueta como nombre común.

Ejemplo:

Forma	***Lema***	***POS***
cm	*centímetro*	*NCMS000*

- **G5. Tratamiento de palabras gramaticales**

a) Tratamiento de *se*

Se es anotado siguiendo las etiquetas que utiliza FreeLing3.1 recogidas en la siguiente tabla. 'P00CN000' se utiliza para los usos de *se* en construcciones impersonales y pasivas reflejas. 'P03CN000' se utiliza cuando *se* forma parte de un predicado pronominal. 'PP3CN000' se utiliza en los casos en los que *se* es un pronombre personal de objeto.

Forma	***Lema***	***POS***
se	*se*	*P00CN000*
se	*se*	*P03CN000*
se	*se*	*PP3CN000*

En caso de error en la anotación automática, la etiqueta se corregirá en el 20% del corpus anotado manualmente.

b) Tratamiento de *que*

Que es anotado siguiendo las etiquetas que utiliza FreeLing3.1 recogidas en la siguiente tabla. Cuando se trata de una conjunción se le asigna el POS 'CS'. Cuando se trata de un pronombre relativo se le asigna el POS 'PR0CN000'.

Forma	***Lema***	***POS***
que	*que*	*PR0CN000*
que	*que*	*CS*

En caso de error en la anotación automática, la etiqueta se corregirá en el 20% del corpus anotado manualmente.

c) Tratamiento de amalgamas

Las amalgamas 'al' y 'del' son etiquetadas según la etiqueta de EAGLES 'SPCMS', tanto en la anotación manual como en la automática.

Ejemplo:

Forma	***Lema***	***POS***
al	*al*	*SPSMS*
del	*del*	*SPSMS*

3.2 REGLAS POSITIVAS (REGLAS-P)

- **P1. Tratamiento de palabras en otros idiomas**

Las palabras que aparecen en otros idiomas se lematizan en el mismo idioma en el que aparecen. En los casos en que algunos atributos del POS no sean identificables, se les asigna el valor 0.

Ejemplo:

Forma	*Lema*	*POS*
analyzer	*analyzer*	*NCCS000*
caecum	*caecum*	*NC00000*

- **P2. Nombres de medicamentos y principios activos en minúscula**

Los nombres de medicamentos y principios activos que se encuentran escritos en minúsculas en el corpus se etiquetan como NC00000. En aquellos casos en los que el género y el número es identificable se incorpora a la etiqueta.

Ejemplo:

Forma	*Lema*	*POS*
diazepam	*diazepam*	*NCMS000*

- **P3. Nombres de medicamentos y principios activos en mayúscula**

Los nombres de medicamentos y principios activos que aparecen con mayúscula en posición no inicial de frase son etiquetados como NP00000.

Ejemplo:

Forma	*Lema*	*POS*
Carbamazepina	*carbamazepina*	*NP00000*

- **P4. Nombres de medicamentos y principios activos en mayúscula y minúscula**

Los nombres de medicamentos y principios activos que aparecen con mayúscula en unos casos y con minúscula en otros en posición no inicial de frase son etiquetados como NC00000.

Ejemplo:

Forma	*Lema*	*POS*

Etanercept	*Etanercept*	*NC00000*
etanercept	*etanercept*	*NC00000*

- **P5. Siglas y abreviaturas**

Las siglas y las abreviaturas son etiquetadas como NC00000, ya que al no desarrollarse el lema (ver regla G3) no es posible obtener en la anotación más datos para los atributos del POS.

Ejemplo:

Forma	***Lema***	***POS***
PAAF	*paaf*	*NC00000*

- **P6. Emails y URLs**

Los emails y las URLs son etiquetadas como NP0000.

Ejemplo:

Forma	***Lema***	***POS***
http://nefrochus.villaweb.es/en/	*http://nefrochus.villaweb.es/en/*	*NP00000*

- **P7. Afijos en estructuras coordinadas**

Los afijos locativos o temporales que aparecen en estructuras coordinadas se anotan como RG.

Ejemplo: *fractura intra o posoperatoria*

Forma	***Lema***	***POS***
intra	*intra*	*RG*

- **P8. Cadenas alfanuméricas**

Las cadenas alfanuméricas que designan compuestos químicos, genes o cromosomas se anotan como NC00000 (ver regla I1).

Ejemplo:

Forma	*Lema*	*POS*
p.R589H	*p.R589H*	*NC00000*

3.3 REGLAS NEGATIVAS (REGLAS-N)

La guía de anotación de POS no contempla reglas negativas ya que la totalidad de los tokens del corpus se anotan con etiqueta morfológica.

3.4 IMPLEMENTACIÓN DE LAS REGLAS EN ANOTACIÓN AUTOMÁTICA (REGLAS-I)

Estas reglas proporcionan los detalles de implementación de las reglas de anotación manual en el proceso de anotación automática de FreeLing3.1.

- **I1. Expresiones alfanuméricas**

El POS de las expresiones alfanuméricas cambia en la anotación automática de Z a NC00000. Para ello se modifica el módulo *User Map Module* de FreeLing que permite alterar el POS de los patrones establecidos dentro del mismo módulo. Esta función se activa mediante el comando:

--usr --fmap fichero-de-patrones

Para asignar el nuevo POS, se establece una expresión regular que engloba las expresiones alfanuméricas contenidas en el corpus a las que se les asigna por defecto la etiqueta NC00000. El lema de dichas expresiones alfanuméricas es la propia forma.

4 BIBLIOGRAFÍA

[1] Barret, N. & Weber-Jahnke, J. (2014) A token centric part-of-speech tagger for biomedical text. *Artificial Intelligence in Medicine*, 61, 11-20.

[2] Campillos, L., Deléger, L., Grouin, C., Hamoon, T., Ligozat, A.L. & Névéol, A. (2018) A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annotated Text corpus (MERLOT). *Language Resources & Evaluation,*

52:571-601.

[3] Fan, J.W., Yang, E.W., Jiang, M., Prasad, R., Loomis, R.M., Zisook, D.S., Denny, J.C., Xu, H. & Huang, Y. (2013) Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. *Journal of the American Medical Informatics Association: JAMIA*, *20*(6), 1168–1177.

[4] Griffis, D., Shivade, C., Fosler-Lussier, E. & Lai, A.M. (2016) A Quantitative Evaluation of Sentence Boundary Detection for the Clinical Domain. *AMIA Summits on Translational Science Proceedings*, 88–97.

[5] He, Ying & Kayaalp, M. (2006) A comparison of 13 Tokenizers on MEDLINE. Technical Report. Available at: https://lhncbc.nlm.nih.gov/publication/lhncbc-tr-2006-003 Access date: 8/06/2018

[6] Kazama, J., Miyao, Y. & Tsujii. J. (2001) A Maximum Entropy Tagger with Unsupervised Hidden Markov Models. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*. November 2001. Tokyo, Japan. 333--340.

[7] Kim, J., Ohta, T. Teteisi, Y. & Tsujii, J. (2006) Genia Corpus Manual: Encoding schemes for the corpus and annotation. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.9947&rep=rep1&type=pdf Access date: 21/06/2018.

[8] Kim J.D., Ohta T., Tateishi Y., & Tsujii J. (2003) GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:suppl. 1):180–i182.

[9] Padró, L & Stanilovsky, E. (2012) FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. May 2012. Istanbul, Turkey. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf Accessed date: 05/07/2018

[10] Pakhomov, S.V., Coden, A. & Chute, C.G. (2006) Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, 75, 418-429.

[11] RAE (2005) Signos ortográficos*. Diccionario panhispánico de dudas Real Academia Española http://lema.rae.es/dpd/srv/search?id=qXGSxldBKD6hqrTMMo*

[12] Sætre, R.,Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y. & Ohta, T. AKANE System: Protein-Protein Interaction Pairs in BioCreAtIvE2 Challenge, PPI-IPS subtask. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. April 2007. 209--212.

[13] Savova, G. K., Masanz, J. J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. & Chute, C. G. (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics*

Association : JAMIA, *17*(5), 507–513. http://doi.org/10.1136/jamia.2009.001560

[14] Teteisi, Y. & Tsujii, J. (2006) Genia Annotation Guidelines for Tokenization and POS Tagging. Available at: http://www.nactem.ac.uk/tsujii/papers/yucca/GENIA_Guidelines_POS.pdf.4 Access date: 21/06/2018

[15] Teteisi, Y. & Tsuji, J. (2004) Part-of-Speech Annotation of Biology Research Abstracts. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004) May 2004, Lisbon, Portugal* 1267-1270. http://www.nactem.ac.uk/aigaion2/index.php?/publications/show/129 Accessed date: 02/07/2018.

[16] Tomanek, K., Wermter, J. & Hahn, U. (2007) Sentence and Token Splitting On Conditional Random Fields. Available at: https://pdfs.semanticscholar.org/5651/b25a78ac8fd5dd65f9c877c67897f58cf817.pdf Access date: 9/06/2018

[17] Warner, C., Lanfranchi, A., O'Gorman, T., Howard, A., Gould, K. & Regan, M. (2012) Bracketing Biomedical Text: An Addendum to Penn Treebank II Guidelines. Available at: https://clear.colorado.edu/compsem/documents/treebank_guidelines.pdf Access date: 11/06/2018

5 GLOSARIO DE SIGLAS Y ACRÓNIMOS

Plan TL	Plan de Impulso de las Tecnologías del Lenguaje
PLN	Procesamiento del Lenguaje Natural
POS	Part of Speech
TA	Traducción Automática