

**Interreg**  
**Sudoe**

European Regional Development Fund



# E2.4 DEVELOPMENT OF SUPERVISED CATEGORIZATION MODELS, TOPIC MODELLING AND EXTRACTION OF CLINICAL INFORMATION

Due date:	30/04/2020
Actual submission date:	30/04/2021
Responsible partner:	BSC
Version:	04
Status:	Final
Dissemination level:	Public / Consortium

<b>Deliverable description:</b>
<p>This deliverable describes the methodology used in WP2 to develop the SUPERVISED CATEGORIZATION MODELS, TOPIC MODELLING AND EXTRACTION OF CLINICAL INFORMATION using deep learning techniques. The results obtained by the deep learning models are remarkable, reaching 91% F1 on average.</p>

<b>Revision history</b>			
<b>Version</b>	<b>Date</b>	<b>Comments</b>	<b>Partner</b>
01	12/2020	First version	BSC
02	02/2021	Second version	BSC
03	04/2021	Final version	BSC

<b>Authors</b>	
<b>Name</b>	<b>Partner</b>
Marta Villegas Montserrat	BSC
Aitor González Agirre	BSC
Joan Llop	BSC
Siamak Barzegar	BSC

<b>Contributors</b>	
<b>Name</b>	<b>Partner</b>

## ABBREVIATIONS AND ACRONYMS

HUSE	Hospital Universitario Son Espases
XML	Extensible Markup Language
HER	Electronic Health Record
TTR	Type Token Ratio
BRAT	Brat Rapid Annotation Tool
F1	F1 score
IAA	Inter-Annotator Agreement
NER	Named Entity Recognition
NERC	Named Entity Recognition and Classification

## TABLE OF CONTENTS

INTRODUCTION .....	6
1. METHODOLOGY .....	6
2. VARIABLES OF INTEREST & MAIN CHALLENGES .....	9
Section headers .....	10
Main diagnosis and related attributes .....	10
Procedures and their attributes .....	11
Treatments .....	13
Rating scales.....	13
3. THE GOLD STANDARD .....	14
4. THE PRE-ANNOTATION TOOL .....	16
5. EVALUATION OF THE RULE BASED PRE-ANNOTATION SYSTEM.....	17
6. DEEP LEARNING METHODS .....	19
Evaluation method.....	21
7. RESULTS .....	24
Biomedical and clinical models.....	27
Summary and conclusions .....	29
Code & Demos.....	30
8. USAGE GUIDELINES.....	32
9. LIST OF FIGURES .....	32
10. LIST OF TABLES.....	33
ANNEX 1 List of non-header variables and their frequency.....	34
ANNEX 2 List of header variables and their frequency.....	36
ANNEX 3 Detailed results for the Biomedical model .....	37
ANNEX 4 Detailed results for the Clinical model .....	40
ANNEX 5 ICTUSnet cTAKES pipeline Installation Guidelines for developers .....	43
ANNEX 6 ICTUSnet cTAKES Developing Guidelines .....	52

## EXECUTIVE SUMMARY

This deliverable describes the methodology used in WP2 to develop the SUPERVISED CATEGORIZATION MODELS, TOPIC MODELLING AND EXTRACTION OF CLINICAL INFORMATION using deep learning techniques. The document fully reports the main challenges of the task, the rule-based system for the pre-annotation task, and the eventual supervised model. The results are reported in detail, and two models are compared. The annexes of the document contain additional material and information. The document contains links to the code in GitHub and the demo developed. **The results of the deep learning models are remarkable, reaching 91% F1 on average and they demonstrate that the use of language technologies can be of great help in clinical information extraction tasks, as in the case of ICTUSnet.**

## INRODUCTION

In this document, we describe the methodology followed when developing the supervised models for clinical information extraction and the results achieved. The objective of the supervised models is to support human experts when identifying and extracting relevant variables from stroke discharge reports to fill in the Ictus Registry. The set of relevant variables (i.e. variables of interest) were defined in WP1. The ultimate objective of the task is to assess the extent to which text mining technologies are able to meet the needs of a scenario such as the one in ICTUSnet. From now on this document is organized as follows:

Section 1 describes the methodology followed in WP2.

Section2 analyzes the variables of interest focusing on the challenges of the clinical information extraction task in the context of the ICTUSnet project.

Section 3 gives some statistical information about the Gold Standard used to train and evaluate the deep learning models.

Section 4 describes the rule based pre-annotation system used to support the manual annotation task. The corresponding installation guidelines and user manual can be found in the Annexes of this document.

Section 5 reports the performance of the rule-based system. In this assessment task, we used the same test set used to evaluate the deep learning models. Note that the objective of this evaluation is just to evaluate the performance of the system, a fair comparison with deep learning models is not possible since the rule based system has already seen the test set.

Section 6 describes the development of the deep learning models and the methodology used to evaluate the systems.

Finally, Section 7 reports and analyses the results, summarizes the main conclusions and gives the links to the demos.

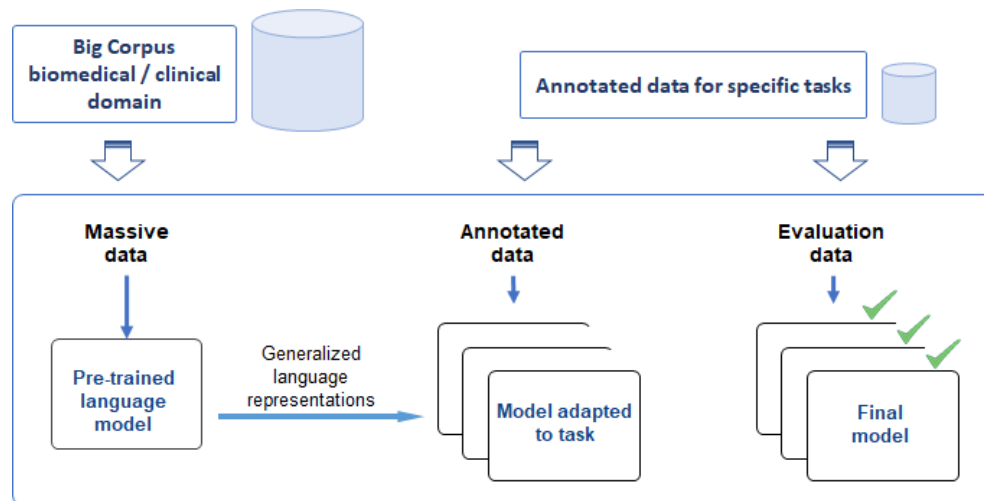
The rest of the document contains list of tables and figures and a number of annexes with supplementary information.

## 1. METHODOLOGY

Following standard deep learning techniques we used neural networks to generate a domain specific language model and, then, we fine-tuned (adapted) the mode to a specific task (i.e. Named Entity Recognition and Classification, NERC).

To train the supervised models for this information extraction task, we need annotated data. The **annotation task** was done by domain experts and governed by annotation guidelines that unambiguously determine the rules to be applied. The **annotation guidelines** used in the project

can be found in the **Deliverable E2.5**. Figure 1 illustrates the process of model generation (using large amounts of biomedical domain data) and model tuning (using a small set of annotated data) to train the model to perform some specific task.



*Figure 1: General schema for model generation and model tuning*

To ease the manual annotation task, we developed a **rule based pre-annotation system** that (i) identifies and normalizes section headers and (ii) identifies and normalizes the variables of interest. For additional information about the section headers’ normalization see **Deliverable E2.3** “Application For The Standardization of Multilingual Clinical Documents”. The automatic pre-annotation system was developed in an iterative way, so that the process was split into different steps, each consisting of 50 to 100 discharge reports. At each new bunch of pre-annotated files, the system was evaluated against the human annotations and modified to improve its performance. For the most part, the improvements consisted of the inclusion of new terms in the dictionary because, as the annotators worked on new reports, they found new variants and forms that initially were not expected. Figure 2 illustrates this iterative process.

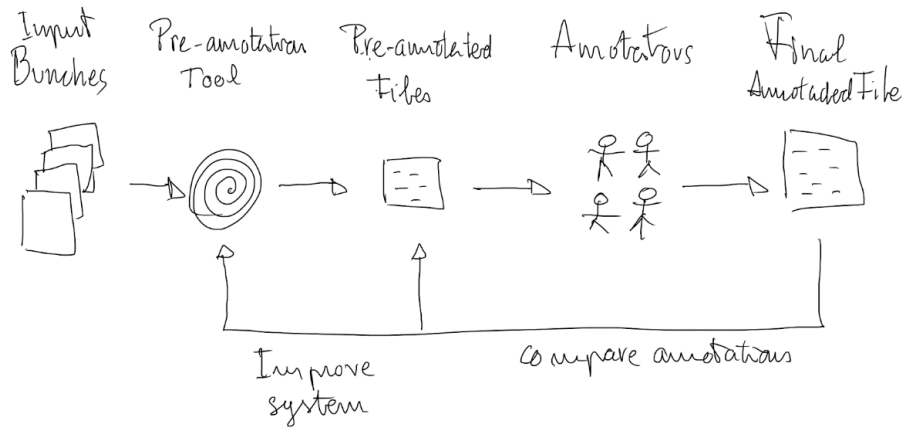


Figure 2 Iterative (pre)-annotation process

For the **manual annotation task** we had a team of 4 annotators (3 nurses and 1 doctor) and used the [BRAT](#) annotation tool, properly configured to our scenario. The training phase was particularly long due to the difficulty of the task and the lack of consensus in certain aspects. The whole manual annotation process included 17 bunches of approximately 50 to 100 files each (some of them were repeated). In the initial training stage, the annotators work together and the guidelines were updated to solve the problems and issues that arose. In the second stage and for the first 5 bunches, all files were annotated by at least two annotators and different **inter-annotator agreement (IAA)** calculus were performed. Once the IAA was good enough, we started the real annotation phase. During the process, we (i) had regular meetings to clarify doubts, (ii) used a WhatsApp group to facilitate communication between annotators and guideline writers and (iii) established a “trouble report” system where annotators collected doubts that were discussed and eventually solved by the responsible of the guidelines. During the whole process, the guidelines were modified and updated accordingly. Figure 3 shows the manual annotation process.

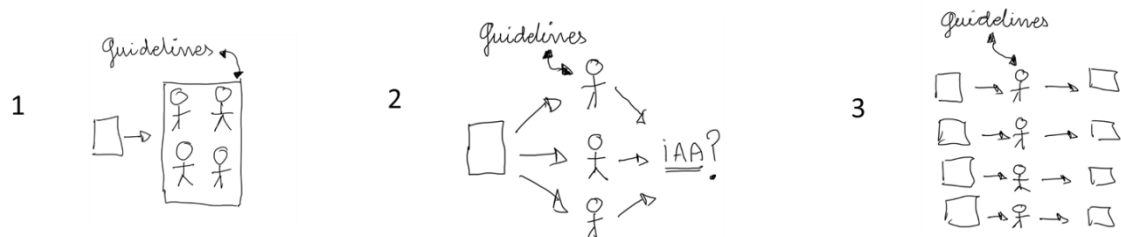


Figure 3 Manual annotation process

As illustrated in Figure 1 above at the beginning of this section, we used a large biomedical data set to generate a **biomedical pre-trained language model**. For this, we collected a **big biomedical corpora** gathering data from a variety of medical resources, namely scientific literature, clinical



cases and crawled data.

We used the resulting corpus to train a **biomedical RoBERTa-base model** with 12 layers/heads and hidden layer size 768, for a total number of 126M parameters.

Then, **we adapted the model to the clinical domain** by overtraining it with 120MB of clinical textual data (including ICTUSnet data provided by AQuAS, Son Espases and IACS). We continued the training process for 48h more, and then selected the best model based on perplexity, using a patience of 20.

Finally, we **fine-tuned our pre-trained models for NER task using the ICTUSnet Gold Standard dataset**. The gold standard was split into train, dev and test sets with standard proportions: 80% for training (656 documents), 10% for valid (83 documents), and 10% for test (83 documents). In this splitting we made sure that the proportions of the diagnoses were preserved for each of the sets. We fine-tuned for 10 epochs and selected the best epoch validating on the dev set.

We used both, the Biomedical and the Clinical models to generate and compare the predictions. Figure 4 illustrates the whole process.

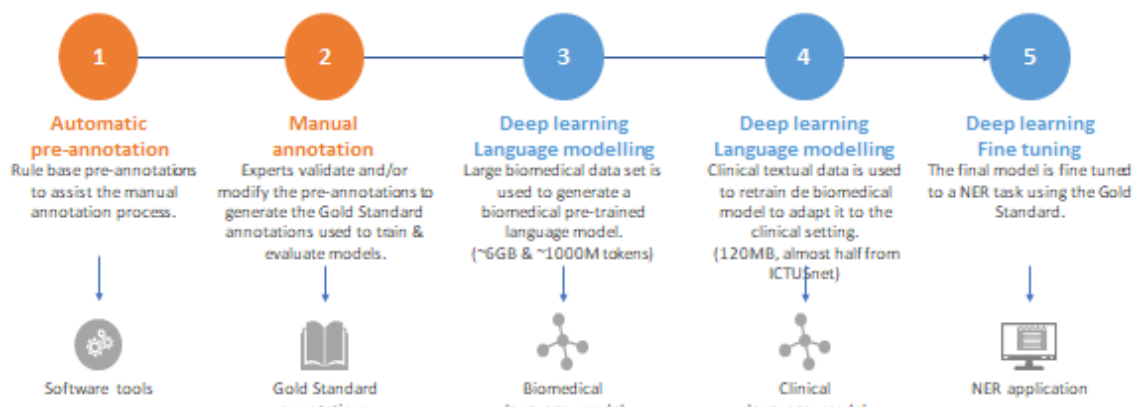


Figure 4: General overview of the methodology

## 2. VARIABLES OF INTEREST & MAIN CHALLENGES

In this section we analyze the kind of variables included in the project and the challenges they pose. As described in the annotations guidelines (see Deliverable 2.5), variables can be classified into four categories as follows.

## Section headers

The objective of the system is to identify and normalize section headers by mapping them into the corresponding **Archetype** (as suggested by the Spanish Ministry). For a detailed description of this normalization process and the pre-annotation tool see **Deliverable E2.3** “Application For The Standardization of Multilingual Clinical Documents”.

Note that section header identification cannot be reduced to a mere dictionary look up task because most elements in the dictionary can only be considered headers under certain circumstances. For example; when they are capitalized and/or follow certain structural patterns. Figure 5 shows some section header annotations in BRAT style: first column has the annotation ID; second column encodes the normalized tag; third and fourth columns serve to encode the initial and final character positions of the mention and, finally, the last column shows the mentions as they occur in the text.

T1	SECCION_MOTIVO_DE_INGRESO	391	408	Motivo de ingreso
T2	SECCION_ANTECEDENTES	475	487	ANTECEDENTES
T3	SECCION_ANTECEDENTES_QUIRURGICOS	880	882	IQ
T4	SECCION_TRATAMIENTO_HABITUAL	909	929	Tratamiento habitual
T5	SECCION_ANTECEDENTES_FAMILIARES	945	968	Antecedentes familiares
T6	SECCION_SITUACION_FUNCIONAL	1148	1172	Situación sociofuncional
T7	SECCION_PROCESO_ACTUAL	1347	1364	ENFERMEDAD ACTUAL
T9	SECCION_EXPLORACION_FISICA	2515	2533	EXPLORACIÓN FÍSICA

Figure 5 Headers' annotations in BRAT style.

## Main diagnosis and related attributes

This includes three main diagnoses: *ictus isquémico*, *ataque isquémico transitorio* and *hemorragia cerebral* and their associated attributes: affected vessel, localization, lateralization and etiology.

The lack of a common naming convention for diagnoses made this task particularly difficult, as often the diagnose is not explicitly named (or it is wrongly named). In our case, for the (pre)-annotation service, diagnosis annotation was addressed as a NER task and an extra diagnosis entity ('other') was added for those underspecified or ambiguous namings that need some kind of interpretation (see the Annotation Guidelines in **Deliverable 3.5** for further details).

Note also that diagnosis and related attributes are '**context dependent**'. As a general rule, the criteria to identify the main diagnosis is by choosing the first disease in the DIAGNOSE section and the rest of diseases in the report are not considered. Similarly, the related attributes are also context dependent. Concretely, they are only relevant provided they are related to the main diagnosis and, consequently, must appear close to it. All other vessels, localizations, lateralizations and etiologies in the text are irrelevant for the task. Figure 6 shows the pre-annotations suggested by the pre-annotation service and Figure 7 shows them in BRAT format.

SECCION DIAGNOSTICOS				
DIAGNÒSTICS				
SUG_Ictus_isquemico	SUG_Arteria_afectada	SUG_Lateralizacion	SUG_Localizacion	SUG_Etiologia
- Ictus isquèmic	de territori de ACM	esquerra	(big lacunar)	de etiologia indeterminada
-Hipertensió arterial essencial				
-Diabetis Mellitus tipus 2				
-Dislipèmia				
-Glaucoma				

Figure 6 Predictions suggested by the automatic pre-annotation tool

```

T21 SECCION_DIAGNOSTICOS 9713 9725 DIAGNÒSTICS
T22 Ictus_isquemico 9728 9735 Infarto
T23 Arteria_afectada 9763 9766 ACM
T24 Lateralizacion 9768 9777 izquierda
T25 Etiologia 9808 9833 etiologia cardioembólico
T27 Arteria_afectada 9885 9887 M2
    
```

Figure 7 Diagnosis annotations in BRAT format

## Procedures and their attributes

This includes five procedures and a number of associated temporal information as listed below:

### Procedures

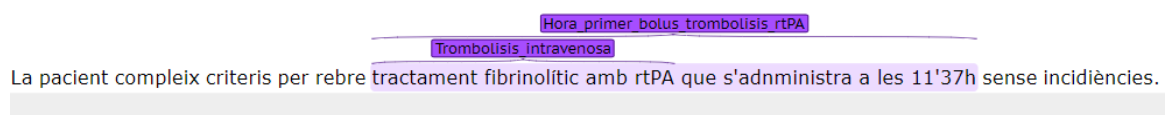
Trombolisis\_intravenosa  
 Trombectomia\_mecanica  
 Trombolisis\_intraarterial  
 Test\_de\_disfagia  
 Tac\_craneal

### Associated temporal information

Fecha Tc craneal inicial  
 Hora Tc craneal inicial  
 Fecha trombólisis iv  
 Hora inicio primer bolus de la trombólisisrtPA  
 Fecha trombectomía mecánica  
 Hora punción arterial para la trombectomía mecánica (groinpuncture)  
 Fecha primera serie para la trombectomía mecánica  
 Hora primera serie para la trombectomía mecánica  
 Fecha recanalización  
 Hora recanalización  
 Fecha finalización trombectomía  
 Hora finalización trombectomía  
 Fecha trombólisis intraarterial

### Hora trombólisis intraarterial

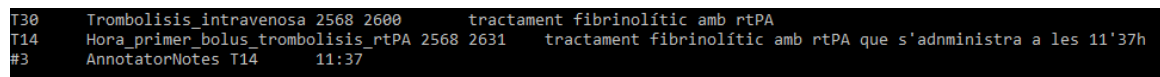
Identifying procedures in the reports is a classical NER task. However, identifying and extracting associated temporal information is a much more complex task. The strategy followed in the guidelines was to ask annotators to annotate (i) mentions of procedures (as in a standard NER task) and (ii) temporal expressions that include a textual part in which the procedure is explicitly mentioned and a temporal expression (a date or time). Figure 8 and Figure 9 show two examples that illustrate the difficulty of the task.



La pacient compleix criteris per rebre tractament fibrinolític amb rtPA que s'administra a les 11'37h sense incidències.

*Figure 8 Annotation of procedures*

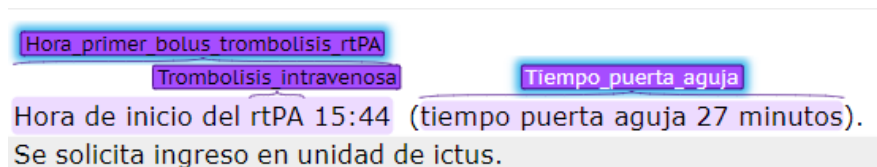
In this example, we have two annotations, one that maps the span *tractament fibrinolític amb rtPA* to the standard form “trombólisis intravenosa”. In the second one, we have a long textual evidence that maps to “hora primer bolus trombólisis rtPA”. In this case, the time information included needs to be identified and normalized. In Figure 9 we can see the same annotation in BRAT style. Note that, for the time variable, we have an extra annotation line where the time expression is normalized.



T30	Trombólisis_intravenosa	2568 2600	tractament fibrinolític amb rtPA
T14	Hora_primer_bolus_trombólisis_rtPA	2568 2631	tractament fibrinolític amb rtPA que s'administra a les 11'37h
#3	AnnotatorNotes	T14	11:37

*Figure 9 Annotations in BRAT style for “Trombólisisintravenosa” and “Hora inicio primer bolus de la trombólisisrtPA”.*

Again, in Figure 10 we have three annotations, one for the Trombólisis\_intravenosa (with textual mention: *rtPA*), another for Hora\_primer\_bolus\_trombólisis\_rtPA (with textual evidence: *Hora de inicio del rtPA 15:44*) and the last one for Tiempo\_puerta\_aguja (with textual evidence: *tiempopuertaaguja 17 minutos*). For the last two examples, we have a textual evidence from where the system needs to extract the relevant temporal information.



Hora primer bolus trombólisis rtPA  
Trombólisis intravenosa Tiempo puerta aguja  
 Hora de inicio del rtPA 15:44 (tiempo puerta aguja 27 minutos).  
 Se solicita ingreso en unidad de ictus.

*Figure 10 Additional annotation examples for procedures*

In these cases, when computing the inter-annotator agreement, only the standardized temporal information (encoded in the BRAT *Notes* field as shown in Figure 9 above) is taken into account.

The span (i.e. the textual evidence) is not evaluated. See Section 6 for more details on the evaluation methods.

For **TAC craneal**, we followed a different strategy: all evidences in text were annotated and related temporal attributes (date and time) were encoded provided they occur in the same line (See the annotation guidelines in Deliverable D2.5 for detailed information). Figure 11 shows the way the annotation of “tac craneal” was addressed: (i) all mentions in text are annotated and (ii) time expressions next to any tac craneal are encoded as associated temporal attributes.

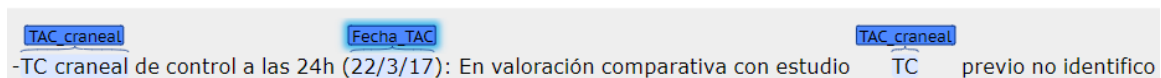


Figure 11 Annotation of TAC craneal and its associated temporal information

## Treatments

For treatments, the objective is to find **anticoagulants** and **antiaggregants** and to classify them as “pre admission medication” or “discharge medication”. This task is essentially a NER task that includes a classification part (pre-admission vs discharge). This classification mostly depends on the context of the mention (i.e. the section in which the medication is listed).

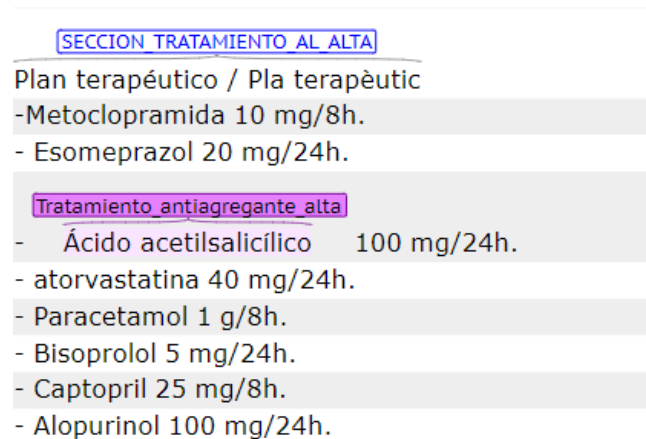


Figure 12 Annotation of treatments

In this example, we can see that *Acido acetilsalicílico* is encoded as “Tratamiento anticoagulante al alta” because it is listed in the section SECCION\_TRATAMIENTO\_AL\_ALTA. Again, these are **context sensitive variables**. All anticoagulants and antiaggregants that fall outside relevant sections are labeled as anticoagulants or antiaggregants without any further classification.

## Rating scales

The relevant scales to be annotated include:

ICTUSnet: E2.4Development of supervised categorization models, topic modelling and extraction of clinical information via cognitive computing.

ASPECTS  
 mRankin\_alta  
 mRankin\_previa  
 NIHSS\_previa  
 NIHSS\_alta

The main challenge in this case is to find the numerical value of the rating scale (often this comes in a complex format, see Figure 13) and to distinguish between *previa/al\_alta* scales. Note, again, that in the vast majority of cases, this *previa/al\_alta* distinction is not explicitly expressed in the reports.

NIHSS\_previa  
 Escala NIHSS (0-2-0-0-2)+(1-0-0-0-0)+(0-1-2-2-0)= 10.

*Figure 13 NIHSS annotation example with a complex numerical sequence*

SECCION ANTECEDENTES QUIRURGICOS  
 IQ:  
 osteosíntesis húmero proximal izq. apendice y colecistectomía. mRankin\_previa  
 mrankin 0.

SECCION SITUACION FUNCIONAL  
 Estado basal: viu amb el marit, autonoma per les AVD.  
mRankin\_previa  
 Rankin previo 2

*Figure 14 mRanking annotation examples*

### 3. THE GOLD STANDARD

The Gold Standard includes a total of 1,006 annotated files with more than 79,000 different annotations. More than 39,000 annotations were **section headers** distributed as follows.

SECTION HEADER	COUNTS
SECCION_PROCESO_ACTUAL	3055
SECCION_EVOLUCION	2904
SECCION_DIAGNOSTICOS	2867
SECCION_EXPLORACION_FISICA	2796
SECCION_EXPLORACIONES_COMPLEMENTARIAS	2772
SECCION_TRATAMIENTO_HABITUAL	2669
SECCION_MOTIVO_DE_INGRESO	2468
SECCION_ANTECEDENTES	2191
SECCION_PROCEDIMIENTOS	1632
SECCION_ANTECEDENTES_PATOLOGICOS	1471
SECCION_TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA	1422
SECCION_TRATAMIENTO_AL_ALTA	1297
SECCION_SITUACION_FUNCIONAL	1272
SECCION_CONTROL	1178
SECCION_DESTINO_AL_ALTA	1136
SECCION_EXPLORACION_FISICA_EN_URGENCIAS	990
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA	937
SECCION_EXPLORACION_FISICA_DURANTE_HOSPITALIZACION	927
SECCION_ANTECEDENTES_PERSONALES	836
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS	778
SECCION_RECOMENDACIONES	738
SECCION_ANTECEDENTES QUIRURGICOS	703
SECCION_EXPLORACION_FISICA_AL_ALTA	452
SECCION_TIPO_DE_INGRESO	444
SECCION_MOTIVO_DEL_ALTA	228
SECCION_ANTECEDENTES_FAMILIARES	205
SECCION_DIAGNOSTICO_PRINCIPAL	179
SECCION_DIAGNOSTICOS_SECUNDARIOS	110
<b>TOTAL</b>	<b>38657</b>

*Table 1 Frequency of section headers*

As reported in **Deliverable D2.2**, section headers show an unbalanced distribution and there are a few frequent headers and along tail of rather infrequent headers. See ANNEX 1 for the list of the rest of the variables, here we just list the 20 top most frequent variables.

VARIABLE	COUNTS
FECHA	11502
HORA	5903
TAC_craneal	4666
Trombolisis_intravenosa	1288
Trombectomia_mecanica	982
NIHSS_previa	964
Fecha_TAC	910
Fecha_de_alta	890
Fecha_de_ingreso	888
Ictus_isquemico	803
Lateralizacion	739
Hora_inicio_sintomas	726
Arteria_afectada	711
Fecha_inicio_sintomas	700
Etiologia	682
mRankin_previa	668
NIHSS	587
Fecha_llegada_hospital	571
NIHSS_alta	570
Tratamiento_antiagregante_alta	550

*Table 2 20 top most frequent variables*

## 4. THE PRE-ANNOTATION TOOL

The pre-annotation system includes three components that are sequentially executed as follows: A **section normalizer**, a python script that identifies and normalizes section headers. The output of the script is an *.ann* file ready to be used in BRAT. See **Deliverable E2.3** “Application For The Standardization of Multilingual Clinical Documents” for further details. The code of this component can be found in GitHub in the following repository: <https://github.com/TeMU-BSC/EHR-HeaderDetector-AnnotationAnalyser>.



An **annotation pipeline developed in cTAKES** framework that identifies and normalizes the variables of interest. The output of the script is an *.ann* file ready to be used in BRAT. See Appendix5 of this document for the documentation about the “**Installation guidelines**” and Appendix6 for the “**Developing guidelines**”. The code of this component can be found in GitHub in the following repository: <https://github.com/TeMU-BSC/spactes>

An **annotation merger** component that: Merges the annotations in (1) and (2) in a single *.ann* file and removes ‘irrelevant’ annotations. (e.g. Removing diagnostic variables such as *Ictus\_isquemico*, *Ataque\_isquemico\_transitorio*, *Hemorragia\_cerebral* and their attributes if they annotated out off the DIAGNOSIS section). The code of the merger component can be found in the following GitHub repository: <https://github.com/TeMU-BSC/brat-merger>

## 5. EVALUATION OF THE RULE BASED PRE-ANNOTATION SYSTEM

We run the rule-based pre-annotation system with the test set defined for the deep learning evaluation to assess the performance of the system and to compare it with the deep learning models (see next section for further details on the train/dev/test split of the gold standard). Note, however, that this is not a fair comparison as, contrary to the deep learning models, the rule-based system already ‘saw’ the test set. This explains, for example, the good performance of the section headers predictions. In this case, for the lexicon look up system, the task was rather easy as all header mentions were in the dictionary. Remember that, in the iterative development approach described in previous section, at each iteration, new mentions are included in the lexicon.

Table 3 reports the results for each variable ordered by frequency, with most frequent variables on top. As we can see in the table, for certain time variables, the results are 0 (marked in red). We decided not to address the annotation of these ‘time variables’ due to the complexity of the task. Annotating this type of information led us to define a list of *ad hoc* regular expressions that was difficult to maintain and did not bring much benefit to the pre-annotation task, so we decided to identify times and dates without going into further classification. Note also that, for treatments and rating scales, we ignored the *previa/alta* distinction and collapsed the two options into a single underspecified tag.

tag	ex	tp	fp	fn	acc	pre	rec	f1
NIHSS	786	726	205	60	0.733	0.780	0.924	0.846
TAC_craneal	625	621	115	4	0.839	0.844	0.994	0.913
mRankin	275	258	16	17	0.887	0.942	0.938	0.940
SECCION_EXPLORACIONES_COMPLEMENTARI AS	217	208	0	9	0.959	1.000	0.959	0.979

SECCION_MOTIVO_DE_INGRESO	203	200	5	3	0.962	0.976	0.985	0.980
SECCION_TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA	200	200	0	0	1.000	1.000	1.000	1.000
SECCION_PROCESO_ACTUAL	185	185	13	0	0.934	0.934	1.000	0.966
SECCION_EXPLORACION_FISICA	177	177	4	0	0.978	0.978	1.000	0.989
SECCION_TRATAMIENTO_HABITUAL	155	149	6	6	0.925	0.961	0.961	0.961
SECCION_EVOLUCION	147	146	0	1	0.993	1.000	0.993	0.997
Trombolisis_intravenosa	146	134	128	12	0.489	0.511	0.918	0.657
Hora_primer_bolus_trombolisis_rtPA	137	0	0	137	0.000	0.000	0.000	0.000
SECCION_ANTECEDENTES	137	133	1	4	0.964	0.993	0.971	0.982
SECCION_EXPLORACION_FISICA_DURANTE_HOSPITALIZACION	136	91	3	45	0.655	0.968	0.669	0.791
Trombectomia_mecanica	122	109	55	13	0.616	0.665	0.893	0.762
SECCION_EXPLORACION_FISICA_EN_URGENCIAS	121	80	0	41	0.661	1.000	0.661	0.796
Ictus_isquemico	117	104	4	13	0.860	0.963	0.889	0.924
SECCION_DESTINO_AL_ALTA	113	109	0	4	0.965	1.000	0.965	0.982
Etiologia	110	88	17	22	0.693	0.838	0.800	0.819
SECCION_DIAGNOSTICOS	110	109	8	1	0.924	0.932	0.991	0.960
ASPECTS	107	106	52	1	0.667	0.671	0.991	0.800
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA	104	96	5	8	0.881	0.950	0.923	0.937
Tratamiento_antiagregante	93	92	42	1	0.681	0.687	0.989	0.811
SECCION_ANTECEDENTES_PATOLOGICOS	91	91	6	0	0.938	0.938	1.000	0.968
SECCION_TRATAMIENTO_AL_ALTA	91	87	3	4	0.926	0.967	0.956	0.961
Tratamiento_anticoagulante	86	86	123	0	0.411	0.411	1.000	0.583
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS	75	70	2	5	0.909	0.972	0.933	0.952
SECCION_SITUACION_FUNCIONAL	72	67	0	5	0.931	1.000	0.931	0.964
Arteria_afectada	64	60	27	4	0.659	0.690	0.938	0.795
Lateralizacion	59	49	8	10	0.731	0.860	0.831	0.845
SECCION_TIPO_DE_INGRESO	53	53	0	0	1.000	1.000	1.000	1.000
Hora_inicio_trombectomia	52	0	0	52	0.000	0.000	0.000	0.000
SECCION_PROCEDIMIENTOS	50	50	1	0	0.980	0.980	1.000	0.990
SECCION_EXPLORACION_FISICA_AL_ALTA	46	46	0	0	1.000	1.000	1.000	1.000
SECCION_ANTECEDENTES_PERSONALES	42	40	0	2	0.952	1.000	0.952	0.976
SECCION_RECOMENDACIONES	42	40	0	2	0.952	1.000	0.952	0.976
Hora_recanalizacion	41	0	0	41	0.000	0.000	0.000	0.000
SECCION_MOTIVO_DEL_ALTA	40	0	0	40	0.000	0.000	0.000	0.000
Hora_primera_serie_trombectomia	36	0	0	36	0.000	0.000	0.000	0.000
Hora_fin_trombectomia	32	0	0	32	0.000	0.000	0.000	0.000

SECCION_ANTECEDENTES_QUIRURGICOS	32	18	4	14	0.500	0.818	0.562	0.667
Localizacion	31	29	35	2	0.439	0.453	0.935	0.611
SECCION_CONTROL	31	31	0	0	1.000	1.000	1.000	1.000
Test_de_disfagia	27	24	0	3	0.889	1.000	0.889	0.941
Tiempo_puerta_aguja	22	0	0	22	0.000	0.000	0.000	0.000
Hemorragia_cerebral	18	17	3	1	0.810	0.850	0.944	0.895
Hora_TAC	17	0	0	17	0.000	0.000	0.000	0.000
SECCION_ANTECEDENTES_FAMILIARES	10	10	0	0	1.000	1.000	1.000	1.000
SECCION_DIAGNOSTICO_PRINCIPAL	10	10	0	0	1.000	1.000	1.000	1.000
SECCION_DIAGNOSTICOS_SECUNDARIOS	9	9	0	0	1.000	1.000	1.000	1.000
Ataque_isquemico_transitorio	8	4	3	4	0.364	0.571	0.500	0.533
<b>ALL</b>	<b>5710</b>	<b>5012</b>	<b>894</b>	<b>698</b>	<b>0.759</b>	<b>0.849</b>	<b>0.878</b>	<b>0.863</b>

*Table 3 Performance of the rule based pre-annotation system. With: number of examples (ex), true positives (tp), false positives (fp), false negatives (fn) accuracy (acc), precision (pre), recall (rec) and F1.*

When considering header sections alone, the performance is much better, reaching an average F1 score of 95%. For diagnosis related variables the system gets 82.21% in F1. The score in this case is lower because diagnosis variables are 'context sensitive' and this poses an additional problem. Note that, when comparing the results with the table above, the precision is much lower compared to recall, as the system produces more false positives. The results demonstrate that context sensitive variables produce false positives.

tag	ex	tp	fp	fn	acc	pre	rec	f1
Arteria_afectada	64	60	27	4	0.659	0.690	0.938	0.795
Ataque_isquemico_transitorio	8	4	3	4	0.364	0.571	0.500	0.533
Etiologia	110	88	17	22	0.693	0.838	0.800	0.819
Hemorragia_cerebral	18	17	3	1	0.810	0.850	0.944	0.895
Ictus_isquemico	117	104	4	13	0.860	0.963	0.889	0.924
Lateralizacion	59	49	8	10	0.731	0.860	0.831	0.845
Localizacion	31	29	35	2	0.439	0.453	0.935	0.611
<b>ALL</b>	<b>407</b>	<b>351</b>	<b>97</b>	<b>56</b>	<b>0.696</b>	<b>0.783</b>	<b>0.862</b>	<b>0.821</b>

*Table 4 Performance of the rule-based pre-annotation system for diagnosis variables*

## 6. DEEP LEARNING METHODS

As introduced in Section 1, we used neural networks to generate a domain specific language model and, then, we adapted the model (fine tuned it) to a specific task (i.e. Named Entity

Recognition, NER). To generate a biomedical language model we need large amounts of biomedical data. We created this **biomedical corpora** gathering data from a variety of medical resources, namely scientific literature, clinical cases and crawled data. We cleaned each corpus independently applying a cleaning pipeline with customized operations designed to read data in different formats, split into sentences, detect the language, remove noisy and bad-formed sentences, finally deduplicate and eventually output the data with their original document boundaries. Finally, in order to avoid repetitive content, we concatenated the entire corpus and deduplicate again between them. Table 5 shows detailed information related to each dataset before and after the cleaning process, in terms of data size, number of sentences and tokens.

<u>Corpus name</u>	<u>Text Size (GB)</u>	<u>Final size (GB)</u>	<u>Raw tokens</u>	<u>Cleaned tokens</u>	<u>Num. sentences</u>
<u>Clinical cases cardiology</u>	0.0035	0,001	149,904.00	147,790.00	9,970.00
<u>Clinical cases radiology</u>	0.0066	0,001	177,366.00	170,997.00	9,948.00
<u>libros_casos_clinicos</u>	0.0083	0,007	1,137,555.00	1,024,797.00	68,833.00
<u>Clinical cases COVID</u>	0.0084	0,001	82,201.00	82,091.00	3,896.00
<u>EMEA corpus</u>	0.087	0,034	13,797,362.00	5,377,448.00	284,575.00
<u>Patents</u>	0.087	0,084	14,022,520.00	13,463,387.00	253,924.00
<u>wikipedia_life_sciences</u>	0.172	0,088	18,771,176.00	13,890,501.00	832,027.00
<u>barr2_background</u>	0.188	0,159	28,868,022.00	24,516,442.00	1,029,600.00
<u>Pubmed</u>	0.211	0,013	1,957,479.00	1,858,966.00	103,674.00
<u>REEC (casos clínicos)</u>	0.823	0,028	4,581,755.00	4,283,453.00	220,726.00
<u>mespen_medline</u>	1.2	0,38	6,864,901.00	4,166,077.00	322,619.00
<u>pdfs_general</u>	3.3		09,124,996.00	7,146,139.00	5,252,481.00
<u>Scielo</u>	3.891	0,631	61,837,972.00	60,007,289.00	2,668,231.00
<u>Medical crawler</u>	606	4,5	?	746,368,185.00	32,766,976.00
<b>TOTAL</b>	<b>615.9858</b>	<b>5,927</b>	<b>261,373,209.00</b>	<b>972,503,562.00</b>	<b>43,827,480.00</b>

*Table 5 The Biomedical corpus*

We used the resulting corpora to train a **Roberta-base model** with 12 layers/heads and a hidden layer sizes of 768 for a total number of 126M parameters. We kept the original Roberta hyperparameter configuration and trained with a masked language model objective. The model was trained for 48 hours using 16 NVIDIA V100 GPUs of 16GB DDRAM. After training, we selected as the best model the checkpoint that achieved the lowest perplexity. Note that, it turned out that the best model for perplexity matched the best model for loss.

Then, we adapted the model to the clinical domain by **further pre-training** with 120MB of clinical textual data (including nearly 34MB of ICTUSnet data provided by AQuAS, Son Espases and IACS). **The data was preprocessed** using a cleaning pipeline with customized operations designed to read data, split documents into sentences, detect the language, and remove noisy and bad-formed sentences. Specifically, the pipeline applies statistical language models, heuristic filters, and hand-written preprocessing rules to keep the documents with the most quality, and restore

or discard sentences that are probably bad-formed or too noisy. Also, the text is formatted such that it can be input to the model. At the end of the process we ended with 151MB of cleaned clinical textual data.

We started from the best model checkpoint obtained on the biomedical corpora and continued the training with two different strategies based on the learning rate, thus generating two models:

- Initializing the learning rate to the same value used at the beginning of the training with biomedical data. We discharged this model based on a preliminary evaluation.
- Use the learning rate value reached by the best checkpoint trained on the biomedical corpora.

We decided to stop the training using an early stopping method on perplexity score with patience of 20 epochs and delta of 0.01 perplexity units.

Finally, to evaluate the resulting two models, we **fine-tuned** them for Named Entity Recognition (NERC) task over the ICTUSNet dataset. The dataset was split into train, dev and test sets with standard 80-10-10 proportions. We fine-tuned for 10 epochs and for each model we selected the best epoch validating on the dev set. For easy of reading, we reproduce here again part of the figure about the methodology in Section 1.

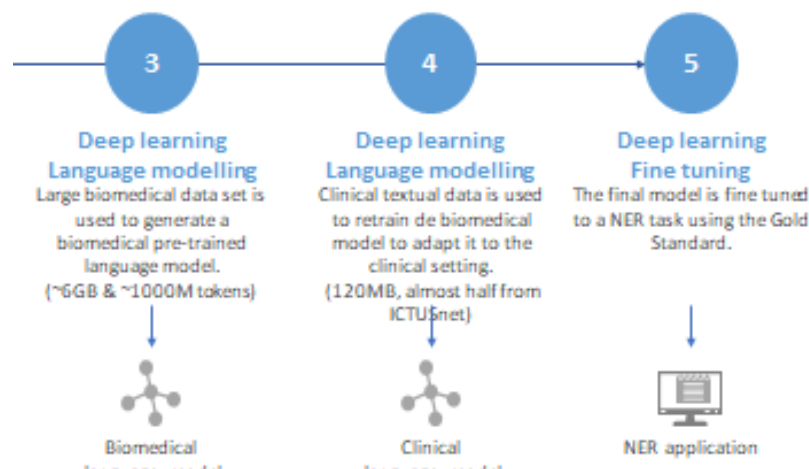


Figure 15 Deep learning process

## The evaluation method

We first split the annotated data (Gold Standard) as follows: 80% for train (656 documents), 10% for valid (83 documents), and 10% for test (83 documents). When splitting, we maintained the same percentage of each of the 3 diagnoses in the splits (ischemic stroke, transient ischemic attack and cerebral hemorrhage). Note that, due to the small number of documents, and the imbalance between these diagnoses, there was a risk that there would be no examples in validation or in test.

Then we moved the annotations from the BRAT standoff format to the BIO/IOB (beginning, inside, outside) format, which is a very common tag format for NER. The prefix "B" in front of a Tag indicates the beginning of a chunk, and an "I" indicates that we are still inside that chunk. The "O" tag is used to indicate that a token does not correspond to any of the entities to be tagged. Table 6 shows an example of a tag phrase in BIO format with one token per line.

token	tag
Vive	O
con	O
su	O
esposa	O
,	O
independiente	O
para	O
ABVD	O
,	O
mRs	B-mRankin_previa
O	I-mRankin_previa
.	O

*Table 6 BIO/IOB format for evaluation*

We used both the Biomedical model and the Clinical model to generate the predictions using the test set and compared them against the correct annotations

Given the high number of tokens assigned to the class "O", we do not take them into account for the case where both for the predictions and the GS we have an O label (this avoids raising the result by the fact that O is the majority class, i.e. that the vast majority of tokens do not belong to any of the entities). Following the previous example in Table 6, only the columns marked in gray in Table 7 would be evaluated (in red text the wrong ones, in green text the correct predictions):

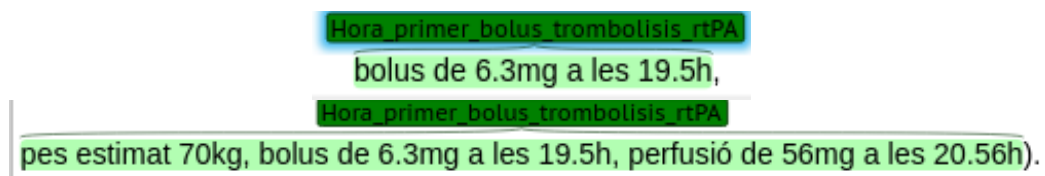
token	GS	prediction
Vive	O	O
con	O	O
su	O	O

esposa	O	O
,	O	O
independente	O	O
para	O	O
ABVD	O	B-Antecedente
,	O	O
mRs	B-mRankin_previa	B-mRankin_previa
O	I-mRankin_previa	I-mRankin_previa
.	O	O

*Table 7 Evaluating annotations*

Once the lines with double O were removed, we evaluated the model using standard metrics: accuracy, precision, recall and F1.

In the case of **dates and times**, the system must provide us with 1) the normalized date or time, and 2) the textual evidence that supports for the previous normalized data or time. As we saw in Section 2, the textual evidences of temporal entities vary greatly in their length, from just the date/time, to full sentences. In this scenario, we only evaluate a prediction as correct if and only if the normalization is exactly the same and the span of the prediction and the human annotation overlap. If the normalization matches, but the spans do not correspond to the same snippet of text, the prediction is considered incorrect. Similarly, if the text spans correspond to the same snippet, but the normalization does not match, the prediction is considered wrong. In the following lines we give some examples of correct/incorrect time predictions. For each example, the first image corresponds to the GS and the second one corresponds to the prediction.



In this example, the spans are clearly different but the tag and the normalized time (19:50) are the same, consequently the annotation is correct.

Hora\_TAC
  
 I (23/02/17; 13:03h) :
   
Hora\_TAC
  
 (23/02/17; 13:03h) :

Again, in this example the two spans are different (note the extra *h* in the second one) but the tag and the normalized time (13:03) are the same. This is a correct prediction.

Hora primer bolus trombolisis rTPA
  
rTPA a las 13.33 h
  
Trombolisis intr Hora primer bolus trombolisis rTPA
  
rTPA a las 13.33 h

Once more, in this new example, the spans are different but the tag and the normalized time expression (13:33) are correct.

Trombolisis intravenosa
  
trombolisis: 80 kg (actilyse bolus 7 mg + perfusión 65 mg en 1 hora)

Trombolisis intravenosa
  
Trombolisis intravenosa
  
Trombolisis intravenosa Hora primer bolus trombolisis rTPA Hora primer bolus trombolisis rTPA
  
trombolisis: 80 kg (actilyse bolus 7 mg + perfusión 65 mg en 1 hora)

Finally, in this example, the prediction clearly fails as it predicts two time tags that are not encoded in the gold standard (image on top).

## 7. RESULTS

Table 8 below shows the initial results for each variable listed by frequency order. As we can see in Figure 16, a good number of variables (15 out of 51) are above 95% in F1 score and almost half of the variables (24 out of 51) are between 76% and 95%. Only 14 variables are below 76% in F1. We marked in red the low results.

Variable	ex	acc	pre	rec	F1
TAC_craneal	621	0.964	0.961	0.984	0.973
NIHSS_previa	383	0.461	0.507	0.687	0.584
NIHSS	238	0.233	0.310	0.338	0.324



SECCION_EXPLORACIONES_COMPLEMENTARIAS	213	0.930	1.000	0.976	0.988
SECCION_MOTIVO_DE_INGRESO	205	0.976	0.965	0.988	0.976
SECCION_TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA	200	0.990	1.000	0.976	0.988
SECCION_PROCESO_ACTUAL	192	0.970	0.935	1.000	0.967
SECCION_EXPLORACION_FISICA	179	0.962	0.963	0.975	0.969
NIHSS_alta	167	0.740	0.855	0.810	0.832
Trombolisis_intravenosa	161	0.824	0.814	0.878	0.845
SECCION_TRATAMIENTO_HABITUAL	155	0.905	0.928	0.987	0.957
mRankin_previa	147	0.801	0.817	0.879	0.847
SECCION_EVOLUCION	147	0.936	0.963	0.987	0.975
SECCION_ANTECEDENTES	138	0.935	0.983	0.922	0.952
Trombectomia_mecanica	119	0.741	0.779	0.815	0.797
Ictus_isquemico	117	0.791	0.841	0.879	0.859
mRankin_alta	117	0.765	0.818	0.766	0.791
SECCION_DIAGNOSTICOS	116	0.894	0.953	0.953	0.953
SECCION_EXPLORACION_FISICA_DURANTE_HOSPITALIZACION	111	0.626	0.667	0.833	0.741
Etiologia	109	0.750	0.833	0.727	0.777
SECCION_DESTINO_AL_ALTA	109	1.000	1.000	1.000	1.000
ASPECTS	107	0.702	0.759	0.820	0.788
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA	102	0.570	0.621	0.692	0.655
SECCION_EXPLORACION_FISICA_EN_URGENCIAS	95	0.735	0.846	0.759	0.800
SECCION_ANTECEDENTES_PATOLOGICOS	93	0.921	0.923	1.000	0.960
SECCION_TRATAMIENTO_AL_ALTA	91	0.905	0.935	0.906	0.921
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS	75	0.867	0.778	0.840	0.808
Arteria_afectada	74	0.659	0.768	0.811	0.789
SECCION_SITUACION_FUNCIONAL	67	0.792	0.857	0.909	0.882
Lateralizacion	53	0.783	0.870	0.887	0.879
SECCION_TIPO_DE_INGRESO	53	1.000	1.000	1.000	1.000
SECCION_EXPLORACION_FISICA_AL_ALTA	51	0.807	0.810	0.895	0.850
SECCION_PROCEDIMIENTOS	51	0.962	0.955	1.000	0.977
Tratamiento_antiagregante_alta	51	0.828	0.886	0.929	0.907
Tratamiento_anticoagulante_alta	48	0.583	0.745	0.729	0.737
SECCION_RECOMENDACIONES	43	0.851	0.909	0.952	0.930
SECCION_ANTECEDENTES_PERSONALES	42	0.909	0.913	0.913	0.913
SECCION_ANTECEDENTES_QUIRURGICOS	31	0.750	0.690	0.909	0.784
SECCION_CONTROL	31	0.879	0.935	0.935	0.935
Localizacion	28	0.614	0.641	0.962	0.769
Test_de_disfagia	27	1.000	1.000	1.000	1.000
Tratamiento_antiagregante_hab	26	0.893	0.926	0.962	0.943

SECCION_MOTIVO_DEL_ALTA	24	0.000	0.000	0.000	0.000
Tratamiento_anticoagulante_hab	22	0.395	0.484	0.682	0.566
Hemorragia_cerebral	18	0.536	0.500	0.727	0.593
Tratamiento_anticoagulante	17	0.292	0.311	0.824	0.452
Tratamiento_antiagregante	15	0.565	0.619	0.867	0.722
SECCION_ANTECEDENTES_FAMILIARES	10	0.833	0.833	1.000	0.909
SECCION_DIAGNOSTICOS_SECUNDARIOS	9	0.889	0.750	0.750	0.750
Ataque_isquemico_transitorio	8	0.333	0.750	0.500	0.600
mRankin	8	0.176	0.286	0.500	0.364
SECCION_DIAGNOSTICO_PRINCIPAL	8	0.364	0.500	0.500	0.500
ALL	5322	0.765	0.833	0.887	0.859

Table 8 Initial results for all variables, Biomedical model

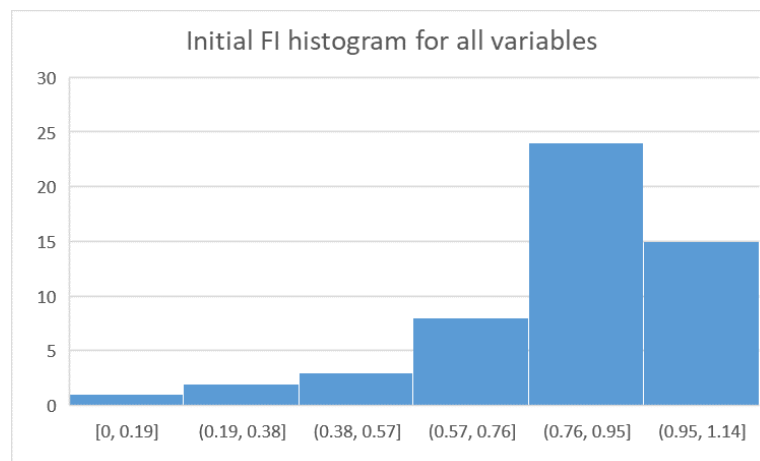


Figure 16 Initial F1 results histogram for the Biomedical model

When analysing the low results in detail, we see that most of them are for low frequency variables and/or have to do with the *previa/alta* distinction discussed in Section 2. Low frequency variables are difficult to assess, models need sufficient examples to learn and we cannot draw any conclusions beyond noting the lack of sufficient data. Especially critical is the case of NHISS variables on top of the table. Here, the system clearly fails to distinguish between *previa/alta*. In most cases, the model correctly assigned the right label in the B(igining)part of the annotation (see Table 7), but assigned a wrong label in the following I(inside) parts of the annotation. Instead of defining a script that forced a label matching between B and I parts of the tags, we decided to collapse tags and ignore this *previa/alta* distinction. As discussed in Section 5, for the task in hand, we considered that suggesting high quality underspecified tags and ask users to classify them was a better strategy than producing low quality pre-annotations. Correcting wrong pre-annotations is a hard task and causes distrust in the system. Consequently, we prepared a new gold standard dataset where these *previa/alta* tags were replaced by underspecified tags and run again the evaluation.

## Biomedical and clinical models

In the following lines we report and compare the eventual results for the Biomedical and Clinical models. See ANNEX 3 and ANNEX 4 further information about the results for each model.

Table 9 shows the global average results comparing both models. As we can see, the differences in true positives and F1 are minimal and show that retraining with clinical data has not brought any advantage to the system. From the results in the table, we can only point out that the Biomedical model has 5.48% more false positives than the Clinical model, and 6.36% less false negatives. In any case the differences in accuracy, precision and recall are insignificant.

Model	Examples	Tp	Fp	Fn	Acc	Pre	Rec	F1
Biomedical	5455	5125	675	330	0.836	0.884	0.940	0.911
Clinical	5455	5104	638	351	0.838	0.889	0.936	0.912

*Table 9 Global average results comparing Biomedical and Clinical models*

In Table 10, we compare the results for all variables. For each variable, we give the frequency, the F1 score in both models and the difference between them (diff column). We highlighted the variables having a difference greater than 3%. As we can see, (i) main differences are for less frequent variables and, (ii) of these, the time variables are better predicted by the clinical model.

tag	examples	Biomedical	Clinical	diff
NIHSS	786	0.847	0.830	1.7
TAC_craneal	625	0.983	0.983	0.0
mRankin	275	0.961	0.966	-0.5
SECCION_EXPLORACIONES_COMPLEMENTARIAS	217	0.954	0.954	0.0
SECCION_MOTIVO_DE_INGRESO	203	0.983	0.974	0.9
SECCION_TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA	200	0.980	0.990	-1.0
SECCION_PROCESO_ACTUAL	185	0.981	0.981	0.0
SECCION_EXPLORACION_FISICA	177	0.969	0.972	-0.3
SECCION_TRATAMIENTO_HABITUAL	155	0.937	0.950	-1.3
SECCION_EVOLUCION	147	0.964	0.961	0.3
Trombolisis_intravenosa	146	0.925	0.906	1.9
SECCION_ANTECEDENTES	137	0.974	0.962	1.2
SECCION_EXPLORACION_FISICA_DURANTE_HOSPITALIZACION	136	0.842	0.803	3.9
Trombectomia_mecanica	122	0.799	0.816	-1.7
SECCION_EXPLORACION_FISICA_EN_URGENCIAS	121	0.886	0.836	5.0
Ictus_isquemico	117	0.895	0.896	-0.1
SECCION_DESTINO_AL_ALTA	113	0.968	0.968	0.0
Etiologia	110	0.854	0.861	-0.7

SECCION_DIAGNOSTICOS	110	0.943	0.960	-1.7
ASPECTS	107	0.811	0.869	-5.8
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA	104	0.755	0.813	-5.8
Tratamiento_antiagregante	93	0.882	0.875	0.7
SECCION_ANTECEDENTES_PATOLOGICOS	91	0.937	0.937	0.0
SECCION_TRATAMIENTO_AL_ALTA	91	0.941	0.945	-0.4
Tratamiento_anticoagulante	86	0.667	0.687	-2.0
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS	75	0.938	0.938	0.0
SECCION_SITUACION_FUNCIONAL	72	0.966	0.973	-0.7
Arteria_afectada	64	0.756	0.764	-0.8
Lateralizacion	59	0.875	0.850	2.5
SECCION_TIPO_DE_INGRESO	53	0.981	0.991	-1.0
SECCION_PROCEDIMIENTOS	50	0.961	0.980	-1.9
SECCION_EXPLORACION_FISICA_AL_ALTA	46	0.893	0.893	0.0
SECCION_ANTECEDENTES_PERSONALES	42	0.988	0.988	0.0
SECCION_RECOMENDACIONES	42	0.864	0.930	-6.6
SECCION_MOTIVO_DEL_ALTA	40	1.000	1.000	0.0
SECCION_ANTECEDENTES_QUIRURGICOS	32	0.853	0.870	-1.7
Localizacion	31	0.703	0.714	-1.1
SECCION_CONTROL	31	0.921	0.984	-6.3
Test_de_disfagia	27	1.000	1.000	0.0
Hora_TAC	22	0.750	0.842	-9.2
Tiempo_puerta_aguja	18	0.955	0.978	-2.3
Hora_primer_bolus_trombolisis_rtPA	18	0.895	0.919	-2.4
Hemorragia_cerebral	17	0.684	0.743	-5.9
Hora_recanalizacion	10	0.667	0.909	-24.2
SECCION_ANTECEDENTES_FAMILIARES	10	0.909	0.824	8.5
SECCION_DIAGNOSTICO_PRINCIPAL	9	0.842	1.000	-15.8
Hora_inicio_trombectomia	9	1.000	0.941	5.9
SECCION_DIAGNOSTICOS_SECUNDARIOS	8	0.941	0.533	40.8
Ataque_isquemico_transitorio	6	0.615	0.769	-15.4
Hora_primera_serie_trombectomia	5	0.600	0.600	0.0
Hora_fin_trombectomia	5	1.000	1.000	0.0
<b>ALL</b>	<b>5455</b>	<b>0.911</b>	<b>0.912</b>	<b>-0.097</b>

*Table 10 Comparing Biomedical and Clinical models*

Figure 17 displays the differences between the models. Blue line shows the frequency of the variables whereas the yellow line shows the differences between the F1 scores in the two models. When the values are positive, the Biomedical model outperforms the clinical one. When the

values are negative, the Clinical model outperforms the biomedical one. The closer to 0, the more equal the results are between the two models. Clearly, the largest differences are found among the less frequent variables.

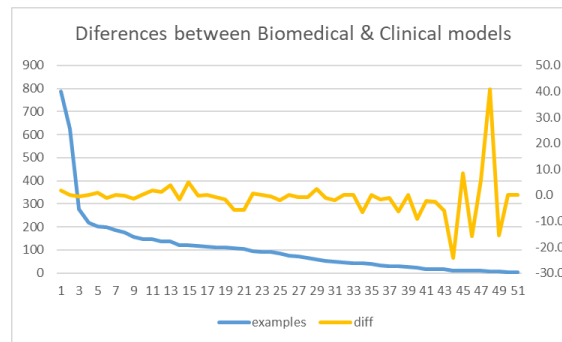


Figure 17 Differences between Biomedical and Clinical models.

## Summary and conclusions

- The information extraction task was complex and ambitious, with 51 different types of variables.
- The manual annotation phase was laborious and long, due to the difficulty of the task and the lack of clear criteria. Only the tremendous effort dedicated by the annotation team, the technological support team, and the experts defining the annotation guidelines made it possible to obtain an eventual gold standard of quality.
- Even so, and from today's perspective, we believe that a better selection of the variables would have yielded better results without detriment to the interest of the project.
- In most cases, the variables were what we have called 'context-dependent', which adds an extra difficulty of the task.
- Temporal variables are a case apart: in most cases the textual evidence shows an enormous variety. Such is the variety that, for the pre-annotation tool, we decided not to address the coding of these variables and limited ourselves to coding dates and times without going any further.
- **The results of the deep learning models are pretty good, reaching 91% F1 on average.** That is without applying any other (post)-process for system improvement. In this exercise we just wanted to evaluate the performance of deep learning techniques.
- The model managed to learn complex aspects such as the 'context sensitivity' (this is very clear in the diagnostic variables, for example).
- The model managed to successfully learn the complex temporal variables that we had given

up in the rule based system. In the following table we can see the good performance despite the low frequencies.

Time variable	Examples	Biomedical	Clinical
Hora_TAC	22	0.750	0.842
Tiempo_puerta_aguja	18	0.955	0.978
Hora_primer_bolus_trombolisis_rtPA	18	0.895	0.919
Hora_recanalizacion	10	0.667	0.909
Hora_inicio_trombectomia	9	1.000	0.941
Hora_primera_serie_trombectomia	5	0.600	0.600
Hora_fin_trombectomia	5	1.000	1.000

*Figure 18 Performance for time variables*

- Retraining with clinical data does not improve the model. We believe that (i) more clinical data (from the stroke domain) would have a better impact and (ii) mixing data from the very beginning would have positive effect, but this remains to be demonstrated
- In any case, we are very satisfied with the results obtained and they demonstrate that the use of language technologies can be of great help in challenging clinical information extraction tasks, as in the case of ICTUSnet.

## Code & Demos

The **code** for the “(Pre)-annotation Pipeline for the ICTUSnet Project” described in this document is dockerised and freely available in the Docker Hub repository (<https://hub.docker.com/r/bsctemu/ictusnet>) to ease its deployment and distribution. The repository contains two different pipelines (tags):

- `bsctemu/ictusnet:ctakes` – with the Initial version based on the Apache cTAKES.
- `bsctemu/ictusnet:deeplearning` – with the Deep Learning version based on transformers.

Since the Docker repository is linked to GitHub (<https://github.com/TeMU-BSC/ictusnet-ctakes>), at any new commit in the GitHub repository, the Docker is automatically updated.

We developed a demo; these are the links to the **demo**, the **demo’s code in GitHub** and the **video tutorial**:

- Link to the demo: <http://temu.bsc.es:81/> (see the screen shoot below)

- Link to GitHub demo's code: <https://github.com/TeMU-BSC/ictusnet-webapp>
- Link to the a video tutorial on YouTube: [https://www.youtube.com/watch?v=uXfAtJK\\_MqA](https://www.youtube.com/watch?v=uXfAtJK_MqA)

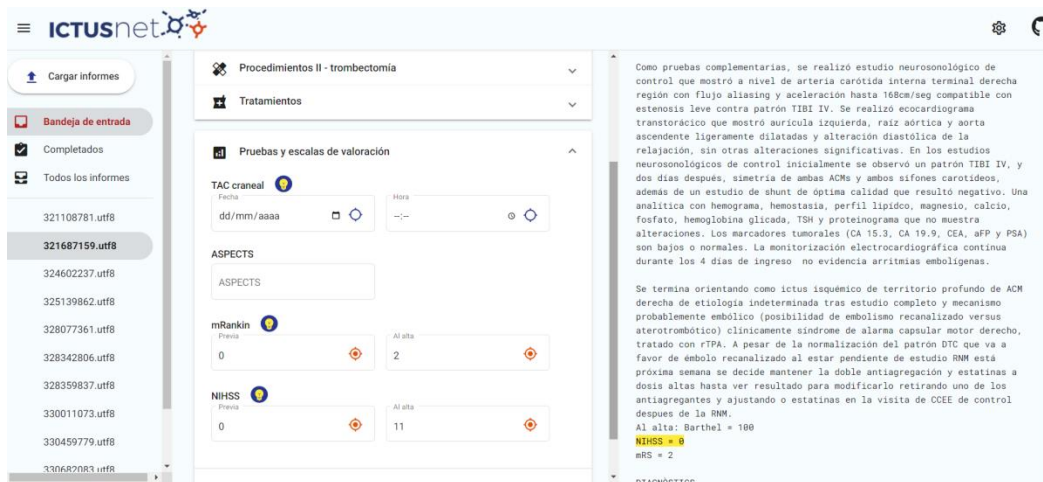


Figure 19 Screenshot of the prototype demo

We also provide a link to **the BRAT annotation tool**, where we can compare the annotations of the gold standard with the predictions made by the deep learning system [https://temu.bsc.es/ICTUSnet/diff.xhtml?diff=%2FICTUSnet time variables and gs%2Ftest brat gs%2F#/ICTUSnet time variables and gs/normalized times test predictions brat/323767062.utf8](https://temu.bsc.es/ICTUSnet/diff.xhtml?diff=%2FICTUSnet+time+variables+and+gs%2Ftest+brat+gs%2F#/ICTUSnet+time+variables+and+gs/normalized+times+test+predictions+brat/323767062.utf8)

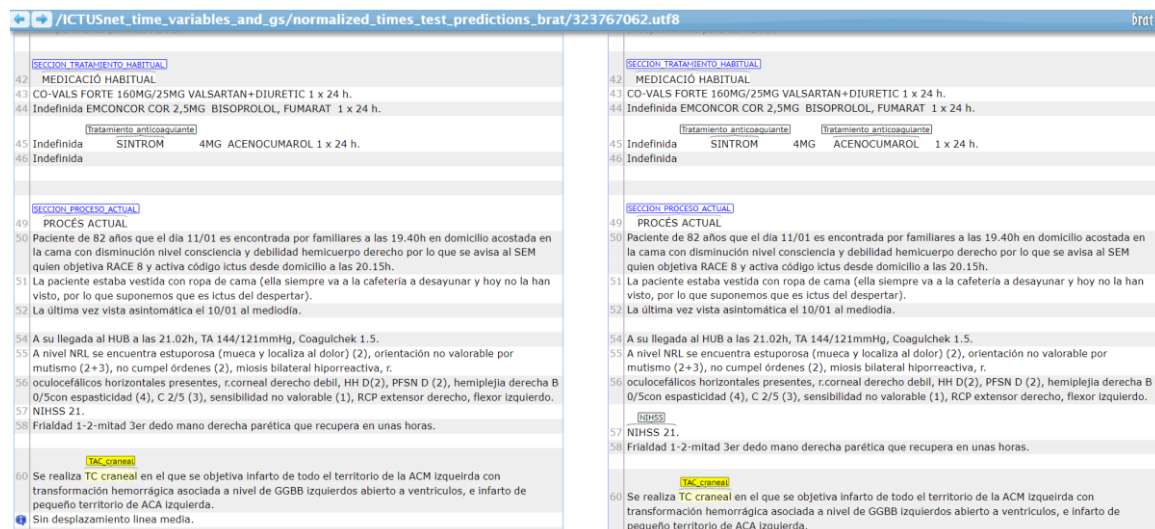


Figure 20 Screenshot of the BRAT tool comparing GS (left) and predictions (right)

## 8. USAGE GUIDELINES

The instructions for the installation and execution of the annotation tool are listed in the readme file of the repository on GitHub <https://github.com/TeMU-BSC/ictusnet-deeplearning>

To see a running example of the system, integrated in a web application, see the live demo tool at <http://temu.bsc.es:81/>. The link to the demo's code is in GitHub <https://github.com/TeMU-BSC/ictusnet-webapp>

The **Deliverable 2.2** gives further details about the demo and how to integrate the system in a web application.

## 9. LIST OF FIGURES

Figure 1: General schema for model generation and model tuning .....	7
Figure 2 Iterative (pre)-annotation process .....	8
Figure 3 Manual annotation process .....	8
Figure 4: General overview of the methodology .....	9
Figure 5 Headers' annotations in BRAT style.....	10
Figure 6 Predictions suggested by the automatic pre-annotation tool .....	11
Figure 7 Diagnosis annotations in BRAT format.....	11
Figure 8 Annotation of procedures .....	12
Figure 9 Annotations in BRAT style for "Trombolisis intravenosa" and "Hora inicio primer bolus de la trombólisis rPA" .....	12
Figure 10 Additional annotation examples for procedures.....	12
Figure 11 Annotation of TAC craneal and its associated temporal information.....	13
Figure 12 Annotation of treatments.....	13
Figure 13 NIHSS annotation example with a complex numerical sequence .....	14
Figure 14 mRanking annotation examples .....	14
Figure 15 Deep learning process .....	21
Figure 16 Initial F1 results histogram for the Biomedical model .....	26
Figure 17 Differences between Biomedical and Clinical models. ....	29
Figure 18 Performance for time variables .....	30
Figure 19 Screenshot of the prototype demo .....	31
Figure 20 Screenshot of the BRAT tool comparing GS (left) and predictions (right) .....	31
Figure 21 Final F1 histogram for the Biomedical model .....	39
Figure 22 F1 histogram for the Clinical model.....	41



## 10. LIST OF TABLES

Table 1 Frequency of section headers .....	15
Table 2 20 top most frequent variables .....	16
Table 3 Performance of the rule based pre-annotation system. With: number of examples (ex), true positives (tp), false positives (fp), false negatives (fn) accuracy (acc), precision (pre), recall (rec) and F1. ....	19
Table 4 Performance of the rule-based pre-annotation system for diagnosis variables. .	19
Table 5 The Biomedical corpus.....	20
Table 6 BIO/IOB format for evaluation .....	22
Table 7 Evaluating annotations.....	23
Table 8 Initial results for all variables, Biomedical model .....	26
Table 9 Global average results comparing Biomedical and Clinical models .....	27
Table 10 Comparing Biomedical and Clinical models.....	28
Table 11 Headers in the whole corpus.....	36
Table 12 Final evaluation of the Biomedical model .....	38
Table 13 Results for diagnosis variables, Biomedical model .....	39
Table 14 Final evaluation for the clinical model.....	41
Table 15 Results for diagnosis variables, Clinical model.....	42

## ANNEX 1 List of non-header variables and their frequency

VARIABLE	COUNT
FECHA	11502
HORA	5903
TAC_craneal	4666
Trombolisis_intravenosa	1288
Trombectomia_mecanica	982
NIHSS_previa	964
Fecha_TAC	910
Fecha_de_alta	890
Fecha_de_ingreso	888
Ictus_isquemico	803
Lateralizacion	739
Hora_inicio_sintomas	726
Arteria_afectada	711
Fecha_inicio_sintomas	700
Etiologia	682
mRankin_previa	668
NIHSS	587
Fecha_llegada_hospital	571
NIHSS_alta	570
Tratamiento_antiagregante_alta	550
Hora_de_alta	506
ASPECTS	479
mRankin_alta	470
Hora_llegada_hospital	466
Localizacion	387
Tratamiento_anticoagulante_alta	376
Tratamiento_antiagregante_hab	330
Hora_TAC	315
Tratamiento_anticoagulante_hab	307
Tratamiento_anticoagulante	233

Hora_primer_bolus_trombolisis_rtPA	226
Tratamiento_antiagregante	151
Hemorragia_cerebral	129
Hora_inicio_trombectomia	127
Test_de_disfagia	113
Hora_recanalizacion	109
Hora_primera_serie_trombectomia	73
Ataque_isquemico_transitorio	70
Hora_fin_trombectomia	65
Tiempo_puerta_aguja	61
Fecha_inicio_trombectomia	54
Fecha_fin_trombectomia	52
Fecha_recanalizacion	49
Fecha_primera_serie_trombectomia	47
mRankin	40
Fecha_trombolisis_rtPA	21
Tiempo_puerta_puncion	9
Trombolisis_intraarterial	3
<b>TOTAL</b>	<b>40568</b>

## ANNEX 2 List of header variables and their frequency

Headers	#Files
PROCESO_ACTUAL	3214
EVOLUCION	3186
EXPLORACION_FISICA	3106
EXPLORACIONES_COMPLEMENTARIAS	3097
DIAGNOSTICOS	3095
TRATAMIENTO_HABITUAL	2937
MOTIVO_DE_INGRESO	2871
ANTECEDENTES	2305
ANTECEDENTES_PATOLOGICOS	1951
PROCEDIMIENTOS	1891
TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA	1871
CONTROL	1446
DESTINO_AL_ALTA	1373
SITUACION_FUNCIONAL	1340
TRATAMIENTO_AL_ALTA	1219
EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA	918
ANTECEDENTES_PERSONALES	878
EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS	764
RECOMENDACIONES	692
EXPLORACION_FISICA_EN_URGENCIAS	637
ANTECEDENTES_QUIRURGICOS	568
TIPO_DE_INGRESO	567
EXPLORACION_FISICA_DURANTE_HOSPITALIZACION	560
EXPLORACION_FISICA_AL_ALTA	552
ANTECEDENTES_FAMILIARES	217
DIAGNOSTICO_PRINCIPAL	196
DIAGNOSTICOS_SECUNDARIOS	129
MOTIVO_DEL_ALTA	15

*Table 11 Headers in the whole corpus*

## ANNEX 3 Detailed results for the Biomedical model

The table below, shows the eventual results for all variables using the Biomedical model. This deep learning model outperforms the rule based system reported in Table 3 in 4.8 points

tag	ex	tp	fp	fn	acc	pre	rec	f1
NIHSS	786	721	195	65	0.735	0.787	0.917	0.847
TAC_craneal	625	618	14	7	0.967	0.978	0.989	0.983
mRankin	275	270	17	5	0.925	0.941	0.982	0.961
SECCION_EXPLORACIONES_COMPLEMETARIAS	217	198	0	19	0.912	1.000	0.912	0.954
SECCION_MOTIVO_DE_INGRESO	203	203	7	0	0.967	0.967	1.000	0.983
SECCION_TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA	200	198	6	2	0.961	0.971	0.990	0.980
SECCION_PROCESO_ACTUAL	185	185	7	0	0.964	0.964	1.000	0.981
SECCION_EXPLORACION_FISICA	177	174	8	3	0.941	0.956	0.983	0.969
SECCION_TRATAMIENTO_HABITUAL	155	150	15	5	0.882	0.909	0.968	0.937
SECCION_EVOLUCION	147	147	11	0	0.930	0.930	1.000	0.964
Trombolisis_intravenosa	146	141	18	5	0.860	0.887	0.966	0.925
SECCION_ANTECEDENTES	137	131	1	6	0.949	0.992	0.956	0.974
SECCION_EXPLORACION_FISICA_DURANTE_HOSPITALIZACION	136	115	22	21	0.728	0.839	0.846	0.842
Trombectomia_mecanica	122	109	42	13	0.665	0.722	0.893	0.799
SECCION_EXPLORACION_FISICA_EN_URGENCIAS	121	105	11	16	0.795	0.905	0.868	0.886
Ictus_isquemico	117	107	15	10	0.811	0.877	0.915	0.895
SECCION_DESTINO_AL_ALTA	113	107	1	6	0.939	0.991	0.947	0.968
Etiologia	110	91	12	19	0.746	0.883	0.827	0.854
SECCION_DIAGNOSTICOS	110	108	11	2	0.893	0.908	0.982	0.943
ASPECTS	107	90	25	17	0.682	0.783	0.841	0.811
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA	104	91	46	13	0.607	0.664	0.875	0.755
Tratamiento_antiagregante	93	86	16	7	0.789	0.843	0.925	0.882
SECCION_ANTECEDENTES_PATOLOGICOS	91	89	10	2	0.881	0.899	0.978	0.937
SECCION_TRATAMIENTO_AL_ALTA	91	87	7	4	0.888	0.926	0.956	0.941
Tratamiento_anticoagulante	86	69	52	17	0.500	0.570	0.802	0.667
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS	75	75	10	0	0.882	0.882	1.000	0.938
SECCION_SITUACION_FUNCIONAL	72	72	5	0	0.935	0.935	1.000	0.966
Arteria_afectada	64	45	10	19	0.608	0.818	0.703	0.756
Lateralizacion	59	49	4	10	0.778	0.925	0.831	0.875

SECCION_TIPO_DE_INGRESO	53	53	2	0	0.964	0.964	1.000	0.981
SECCION_PROCEDIMIENTOS	50	49	3	1	0.925	0.942	0.980	0.961
SECCION_EXPLORACION_FISICA_AL_ALTA	46	46	11	0	0.807	0.807	1.000	0.893
SECCION_ANTECEDENTES_PERSONALES	42	42	1	0	0.977	0.977	1.000	0.988
SECCION_RECOMENDACIONES	42	35	4	7	0.761	0.897	0.833	0.864
SECCION_MOTIVO_DEL_ALTA	40	40	0	0	1.000	1.000	1.000	1.000
SECCION_ANTECEDENTES QUIRURGICOS	32	29	7	3	0.744	0.806	0.906	0.853
Localizacion	31	26	17	5	0.542	0.605	0.839	0.703
SECCION_CONTROL	31	29	3	2	0.853	0.906	0.935	0.921
Test_de_disfagia	27	27	0	0	1.000	1.000	1.000	1.000
Hora_TAC	22	21	1	1	0.600	0.682	0.833	0.750
Tiempo_puerta_aguja	18	13	7	5	0.913	0.955	0.955	0.955
Hora_primer_bolus_trombolisis_rtPA	18	15	7	3	0.810	0.810	1.000	0.895
Hemorragia_cerebral	17	17	4	0	0.520	0.650	0.722	0.684
Hora_recanalizacion	10	10	2	0	0.500	0.556	0.833	0.667
SECCION_ANTECEDENTES_FAMILIARES	10	8	1	2	0.833	0.833	1.000	0.909
SECCION_DIAGNOSTICO_PRINCIPAL	9	9	0	0	0.727	0.889	0.800	0.842
Hora_inicio_trombectomia	9	8	0	1	1.000	1.000	1.000	1.000
SECCION_DIAGNOSTICOS_SECUNDARIOS	8	4	1	4	0.889	1.000	0.889	0.941
Ataque_isquemico_transitorio	6	5	4	1	0.444	0.800	0.500	0.615
Hora_primera_serie_trombectomia	5	5	0	0	0.429	0.600	0.600	0.600
Hora_fin_trombectomia	5	3	2	2	1.000	1.000	1.000	1.000
ALL	5455	5125	675	330	0.836	0.884	0.940	<b>0.911</b>

*Table 12 Final evaluation of the Biomedical model*

Figure 21 shows the histogram for the F1 scores. We can see that nearly half of the variables (25 out of 51) get a score of 93.33% or higher.

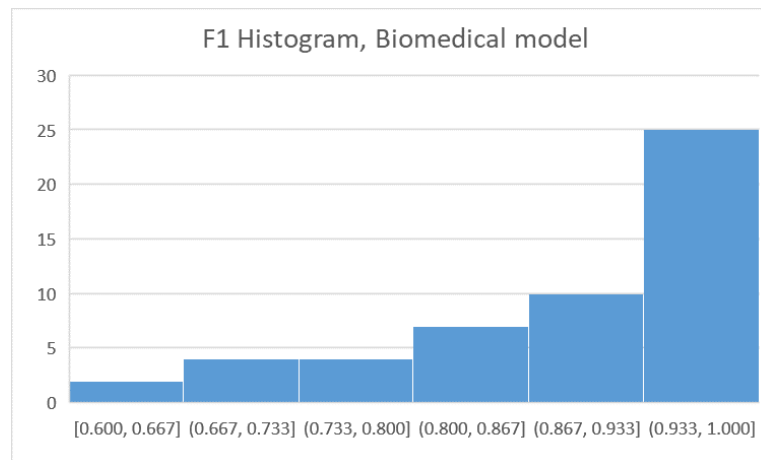


Figure 21 Final F1 histogram for the Biomedical model

When considering header sections alone, the performance reaches an average F1 score of 95.7%. For diagnosis related variables the system gets a 82.3% in F1 (Table 13 **Error! Reference source not found.**). These results are very similar to those of the rule-base system.

tag	ex	tp	fp	fn	acc	pre	rec	f1
Arteria_afectada	64	45	10	19	0.608	0.818	0.703	0.756
Ataque_isquemico_transitorio	8	4	1	4	0.444	0.800	0.500	0.615
Etiologia	110	91	12	19	0.746	0.883	0.827	0.854
Hemorragia_cerebral	18	13	7	5	0.520	0.650	0.722	0.684
Ictus_isquemico	117	107	15	10	0.811	0.877	0.915	0.895
Lateralizacion	59	49	4	10	0.778	0.925	0.831	0.875
Localizacion	31	26	17	5	0.542	0.605	0.839	0.703
<b>ALL</b>	<b>407</b>	<b>335</b>	<b>66</b>	<b>72</b>	<b>0.708</b>	<b>0.835</b>	<b>0.823</b>	<b>0.829</b>

Table 13 Results for diagnosis variables, Biomedical model

## ANNEX 4 Detailed results for the Clinical model

Table 14 shows the eventual results for all variables using the Clinical model.

tag	ex	tp	fp	fn	acc	pre	rec	f1
NIHSS	786	694	192	92	0.710	0.783	0.883	0.830
TAC_craneal	625	615	11	10	0.967	0.982	0.984	0.983
mRankin	275	271	15	4	0.934	0.948	0.985	0.966
SECCION_EXPLORACIONES_COMPLEMENTARIAS	217	198	0	19	0.912	1.000	0.912	0.954
SECCION_MOTIVO_DE_INGRESO	203	203	11	0	0.949	0.949	1.000	0.974
SECCION_TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA	200	198	2	2	0.980	0.990	0.990	0.990
SECCION_PROCESO_ACTUAL	185	183	5	2	0.963	0.973	0.989	0.981
SECCION_EXPLORACION_FISICA	177	174	7	3	0.946	0.961	0.983	0.972
SECCION_TRATAMIENTO_HABITUAL	155	153	14	2	0.905	0.916	0.987	0.950
SECCION_EVOLUCION	147	147	12	0	0.925	0.925	1.000	0.961
Trombolisis_intravenosa	146	139	22	7	0.827	0.863	0.952	0.906
SECCION_ANTECEDENTES	137	128	1	9	0.928	0.992	0.934	0.962
SECCION_EXPLORACION_FISICA_DURANTE_HOSPITALIZACION	136	116	37	20	0.671	0.758	0.853	0.803
Trombectomia_mecanica	122	109	36	13	0.690	0.752	0.893	0.816
SECCION_EXPLORACION_FISICA_EN_URGENCIAS	121	94	10	27	0.718	0.904	0.777	0.836
Ictus_isquemico	117	108	16	9	0.812	0.871	0.923	0.896
SECCION_DESTINO_AL_ALTA	113	106	0	7	0.938	1.000	0.938	0.968
Etiologia	110	93	13	17	0.756	0.877	0.845	0.861
SECCION_DIAGNOSTICOS	110	107	6	3	0.922	0.947	0.973	0.960
ASPECTS	107	96	18	11	0.768	0.842	0.897	0.869
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA	104	100	42	4	0.685	0.704	0.962	0.813
Tratamiento_antiagregante	93	84	15	9	0.778	0.848	0.903	0.875
SECCION_ANTECEDENTES_PATOLOGICOS	91	89	10	2	0.881	0.899	0.978	0.937
SECCION_TRATAMIENTO_AL_ALTA	91	86	5	5	0.896	0.945	0.945	0.945
Tratamiento_anticoagulante	86	68	44	18	0.523	0.607	0.791	0.687
SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS	75	75	10	0	0.882	0.882	1.000	0.938
SECCION_SITUACION_FUNCIONAL	72	72	4	0	0.947	0.947	1.000	0.973
Arteria_afectada	64	47	12	17	0.618	0.797	0.734	0.764
Lateralizacion	59	48	6	11	0.738	0.889	0.814	0.850
SECCION_TIPO_DE_INGRESO	53	53	1	0	0.981	0.981	1.000	0.991
SECCION_PROCEDIMIENTOS	50	50	2	0	0.962	0.962	1.000	0.980



SECCION_EXPLORACION_FISICA_AL_ALT A	46	46	11	0	0.807	0.807	1.000	0.893
SECCION_ANTECEDENTES_PERSONALES	42	42	1	0	0.977	0.977	1.000	0.988
SECCION_RECOMENDACIONES	42	40	4	2	0.870	0.909	0.952	0.930
SECCION_MOTIVO_DEL_ALTA	40	40	0	0	1.000	1.000	1.000	1.000
SECCION_ANTECEDENTES_QUIRURGICOS	32	30	7	2	0.769	0.811	0.938	0.870
Localizacion	31	25	14	6	0.556	0.641	0.806	0.714
SECCION_CONTROL	31	31	1	0	0.969	0.969	1.000	0.984
Test_de_disfagia	27	27	0	0	1.000	1.000	1.000	1.000
Hora_TAC	22	21	1	1	0.727	0.800	0.889	0.842
Tiempo_puerta_aguja	18	13	7	5	0.957	0.957	1.000	0.978
Hora_primer_bolus_trombolisis_rtPA	18	15	7	3	0.850	0.850	1.000	0.919
Hemorragia_cerebral	17	17	4	0	0.591	0.765	0.722	0.743
SECCION_ANTECEDENTES_FAMILIARES	10	10	2	0	0.833	0.833	1.000	0.909
SECCION_DIAGNOSTICO_PRINCIPAL	10	8	1	2	0.700	1.000	0.700	0.824
Hora_inicio_trombectomia	9	9	0	0	1.000	1.000	1.000	1.000
SECCION_DIAGNOSTICOS_SECUNDARIOS	9	8	0	1	0.889	1.000	0.889	0.941
Ataque_isquemico_transitorio	8	4	1	4	0.364	0.571	0.500	0.533
Hora_recanalizacion	6	5	4	1	0.625	0.714	0.833	0.769
Hora_primera_serie_trombectomia	5	5	0	0	0.429	0.600	0.600	0.600
Hora_fin_trombectomia	5	3	2	2	1.000	1.000	1.000	1.000
<b>ALL</b>	<b>5455</b>	<b>5125</b>	<b>675</b>	<b>330</b>	<b>0.838</b>	<b>0.889</b>	<b>0.936</b>	<b>0.912</b>

Table 14 Final evaluation for the clinical model

Figure 22 shows the histogram for the F1 scores.

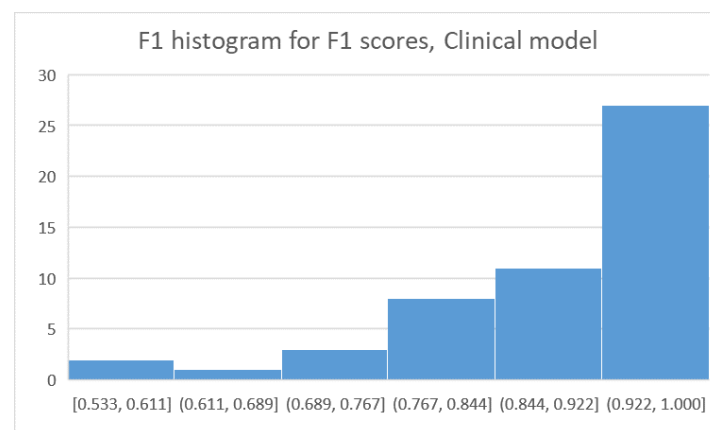


Figure 22 F1 histogram for the Clinical model

When considering header sections alone, the performance reaches and average F1 score of 94%. For diagnosis related variables the system gets a 83.1% in F1 (Table 15 Results for diagnosis variables, Clinical model). These results are very similar to those of the rule-base system.

tag	ex	tp	fp	fn	acc	pre	rec	f1
Arteria_afectada	64	47	12	17	0.618	0.797	0.734	0.764
Ataque_isquemico_transitorio	8	4	3	4	0.364	0.571	0.500	0.533
Etiologia	110	93	13	17	0.756	0.877	0.845	0.861
Hemorragia_cerebral	18	13	4	5	0.591	0.765	0.722	0.743
Ictus_isquemico	117	108	16	9	0.812	0.871	0.923	0.896
Lateralizacion	59	48	6	11	0.738	0.889	0.814	0.850
Localizacion	31	25	14	6	0.556	0.641	0.806	0.714
<b>ALL</b>	<b>407</b>	<b>338</b>	<b>68</b>	<b>69</b>	<b>0.712</b>	<b>0.833</b>	<b>0.830</b>	<b>0.831</b>

*Table 15 Results for diagnosis variables, Clinical model*

## ANNEX 5 ICTUSnet cTAKES pipeline Installation Guidelines for developers

- Introduction
- Install Ubuntu on VirtualBox
- Install FreeLing 4.1
- Download cTAKES
- Download and Install ICTUSnet-cTAKES
- RUN SpaCTeS-cTAKES pipeline for ICTUSnet
- Download Eclipse
- RUN Pipeline by Eclipse
- Install Docker
- Update Docker Repository & Build Docker Image
- Push Image to Docker Hub
- Run Docker Image
- Build Singularity Image
- Run Singularity Image
- Pushing Changes to SpaCTeS repository

### Introduction

This document gives detailed instructions for the installation of the ICTUSnet pipeline and required software.

Section 2 give instructions on how to install a Ubuntu VirtualBox in case you do not have a linux OS

Section 3 describes Freeling installation. Freeling is required by different components of the pipeline and needs to be installed.

Section 4 to 6 include the guidelines for downloading cTAKES, adding ICTUSnet components into the framework, installing ICTUSnet-cTAKES using Maven and, finally, running the pipeline from the command line.

Section 7 and 8 describe how to install Eclipse, import the ICTUSnet-cTAKES project into the IDE and how to run the pipeline.

**We already have a docker image of the current ICTUSnet-cTAKES pipeline** that can be accessed from [Docker Hub](#) or, alternatively you can download the Docker files from [Github Repository](#)

In case you modify the source code of ICTUSnet-cTAKES (by adding components and/or modifying them):

1. you have to build a new image following the instructions in sections 9 to 12 and
2. push the modification code into the [Github Repository](#) as described in section 15.

Sections 13 and 14 show how to build and run a Singularity image in case you need to run the pipeline in Nord3 environment.

### Install Ubuntu on VirtualBox

If you do not have a fresh VM or linux OS you need to install Ubuntu 18.04 (ubuntu-18.04.4-desktop-amd64.iso) on VirtualBox. Note you will need at least 60G HD, and 8 GB Ram

- **VirtualBox (6.1)** — A virtualizer that runs virtual machines
- **Ubuntu OS** — The Linux OS we'll be running in VirtualBox

### Install FreeLing 4.1

```
sudo apt-get -qq update
sudo apt-get -qq upgrade
sudo apt-get install -qqy software-properties-common curl git
build-essential --fix-missing
sudo add-apt-repository ppa:linuxuprising/java
sudo apt update
sudo apt install oracle-java15-installer
sudo apt-get -qqy install language-pack-en-base
sudo update-locale LANG=en_US.UTF-8
sudo nano /etc/default/locale
# add this two variables at end of the file
LANGUAGE="en_US.UTF-8"
LC_ALL="en_US.UTF-8"
sudo locale-gen en_US.UTF-8
sudo apt-get install -y automakeautoconflibtoolwget swig build-
essential
sudo apt-get install -y libboost-regex-dev libicu-dev zlib1g-dev
libboost-system-dev libboost-program-options-dev libboost-thread-
dev libboost-filesystem-dev maven
sudo apt-get install libboost-iostreams-dev
# get cmake.. because freeing needs version above 3.19 and ubuntu installs 3.5.1 (it is
recommend to check the Freeling requirements)
get https://cmake.org/files/v3.19/cmake-3.19.0-Linux-x86_64.sh
sudomkdir /opt/cmake
sudosh cmake-3.19.0-Linux-x86_64.sh --prefix=/opt/cmake --skip-
license
sudo ln -s /opt/cmake/bin/cmake /usr/local/bin/cmake
```

```
# install java8
```

```
sudo apt install openjdk-8-jdk openjdk-8-jre
```

**OPTIONAL:**

```
# It seems Freeling work correctly just with java 8, so if the current version is not 8, change it to the 8 by
```

```
sudo update-java-alternatives --list
```

```
#to list off all the Java installations on a machine by name and directory, and then run
```

```
#sudo update-java-alternatives --set [JDK/JRE name e.g. java-1.8.0-openjdk-amd64]
```

```
sudo update-java-alternatives --set java-1.8.0-openjdk-amd64
```

```
skamler@skamler-VirtualBox:~/Documents$ sudo update-java-alternatives --set java-1.8.0-openjdk-amd64
update-alternatives: error: no alternatives for jaotc
update-alternatives: error: no alternatives for jdeprscan
update-alternatives: error: no alternatives for jhsdb
update-alternatives: error: no alternatives for jimage
update-alternatives: error: no alternatives for jlink
update-alternatives: error: no alternatives for jmod
update-alternatives: error: no alternatives for jshell
update-alternatives: error: no alternatives for mozilla-javaplugin.so
update-java-alternatives: plugin alternative does not exist: /usr/lib/jvm/java-8-openjdk-amd64/jre/lib/amd64/IcedTeaPlugin.so
```

```
# to choose which JRE/JDK to use.
```

```
#If you want to use different JDKs/JREs for each Java task, you can run update-alternatives to configure one java executable at a time; you can run
```

```
sudo update-alternatives --config java
```

```
#freeLing needs this info for installation, so we don't need to keep them or store in /etc/profile
```

```
JAVADIR=/usr/lib/jvm/java-8-openjdk-amd64
```

```
SWIGDIR=/usr/share/swig3.0
```

```
FREELINGDIR=/usr/local
```

```
FREELINGOUT=/usr/local/lib
```

```
export LD_LIBRARY_PATH=/usr/local/share/freeling/APIs/java/
```

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

```
# Setup JAVA_HOME and PATH
```

```
sudo nano /etc/profile
```

```
# add this lines into last line:
```

```
JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

```
export PATH=$PATH:$JAVA_HOME/bin
```

```
export JAVA_HOME
```

```
source /etc/profile
```

```
# Install Freeling from github (2.09 GB) --- take a cofee in the meanwhile
```

```
git clone --depth=1 https://github.com/TALP-UPC/FreeLing.git
```

```
cd FreeLing
mkdir build
cd build
# If we want to have a python wrapper of Freeling as well follow these steps:
sudo apt-get install python3-dev python3-pip python3-tk python3-
lxml python3-six
cmake -DJAVA_API=ON -DPYTHON3_API=on ..
#otherwise
#cmake -DJAVA_API=ON ..
sudo make install
#remove extra files:
sudo apt-get autoremove -y
sudo apt-get clean -y
sudorm -rf /var/lib/apt/lists/* /tmp/* /var/tmp/*
# Now we should have access to the Jfreeling jar wrapper
# This jar file should added to eclipse as external library for Freeling wrapper component
ls -la /usr/local/share/freeling/APIs/java/Jfreeling.jar
```

### **Download cTAKES**

```
sudo apt-get install git
mkdir -p $HOME/Documents # If Documents dir is not exist
cd $HOME/Documents
# Download takes a lot of time (6.24 GB ) --- take a coffee in the meanwhile
# You should have a good connection to internet
git clone https://github.com/apache/ctakes
# or
git clone git@github.com:apache/ctakes.git
```

### **Download and Install ICTUSnet-cTAKES**

```
cd $HOME/Documents
git clone https://github.com/TeMU-BSC/spactes.gitSpaCTeS# Around 54
MB)
cp -rf $HOME/Documents/SpaCTeS/* $HOME/Documents/ctakes/
sudorm -rf $HOME/Documents/SpaCTeS
cd $HOME/Documents/ctakes
mvn clean install -Dgpg.skip -Dmaven.test.skip=true -
DskipTests=false
cd $HOME/Documents/ctakes/ctakes-SpaCTeS
```

```
# get all dependency library for SpaCTeS-cTAKES project
mvn -DoutputDirectory=$HOME/Documents/ctakes/ctakes-
SpaCTeS/target/lib dependency:copy-dependencies
mkdir -p $HOME/Documents/ctakes/ctakes-SpaCTeS/target/config/
cp $HOME/Documents/ctakes/ctakes-core-res/target/classes/log4j.xml
$HOME/Documents/ctakes/ctakes-SpaCTeS/target/config/log4j.xml
```

## **RUN SpaCTeS-cTAKES pipeline for ICTUSnet**

```
# Just for test (In input directory we should have only .txt files)
sudomkdir -p $HOME/Documents/data/TXT
# All txt files should be added to this directory
# Output would be in :$HOME/Documents/data/OUTPUT
sudo java -Djava.library.path=/usr/local/share/freeling/APIs/java
-cp $HOME/Documents/ctakes/ctakes-
SpaCTeS/target/lib/*:$HOME/Documents/ctakes/ctakes-
SpaCTeS/target/ctakes-SpaCTeS-4.0.1-SNAPSHOT.jar -
Dlog4j.configuration=file:$HOME/Documents/ctakes/ctakes-
SpaCTeS/target/config/log4j.xml -Xmx8g
org.apache.ctakes.spactes.pipeline.SpaCTeSBuilderRunner
$HOME/Documents/data/TXT $HOME/Documents/data/OUTPUT
```

## **Download [Eclipse](#)**

```
# Recommended Just for Developing
# Eclipse IDE for Java Developers Version": 2020-03 (4.15.0)
cd $HOME/Documents
wgethttps://mirrors.dotsrc.org/eclipse//technology/epp/downloads/r
elease/2020-03/R/eclipse-java-2020-03-R-linux-gtk-x86\_64.tar.gz
tar -xvf eclipse-java-2020-03-R-linux-gtk-x86_64.tar.gz
rm -rf eclipse-java-2020-03-R-linux-gtk-x86_64.tar.gz
```

## **RUN Pipeline by Eclipse**

```
# Recommended Just for Developing mode
cd $HOME/Documents/eclipse/
./eclipse
# From eclipse do these steps to import ictusnet-ctakes codes
From the file menu, click on import...
Maven -> Existing Maven Projects, then click on Next button
In "Root Directory", brows -> ~/Documents/ctakes, Select all
subprojects, then click on Finish button.
# If there is any error follow following steps for solving:
Right click on ctakes-SPaCTeS -> Maven -> Update project
```

In new window, select all suggested projects (Add out-of-date), then click on update button

From Project menu, click clean button

In the new window, click clean all projects and also build all projects

*# This jar file should added to eclipse as external library for Freeling wrapper component (Jfreeling.jar)*

- Right click on ctakes-openmited-freeling
- Click on properties
- Click on Java Build Path
- Click on Libraries
- Edit Native library location and add this bellow path to it:

```
/usr/local/share/freeling/APIs/java/
```

*# Run in Eclipse*

```
From ctakes-SpaCTeS -> to src/main.src ->  
>org.apache.ctakes.spactes.pipeline
```

```
Open SpaCTeSBuilderRunner.java
```

```
Change INPUT_DIR and OUTPUT_DIR pipelines by adding following path  
Arguments of the SpaCTeSBuilderRunner class (for example:)
```

```
/home/Documents/data/input /home/Documents/data/output
```

```
Right click on SpaCTeSBuilderRunner.java and click on RUN AS ->  
Java Application
```

## Install Docker

```
sudo apt-get update  
sudo apt install docker.io  
sudo systemctl start docker  
sudo systemctl enable docker
```

## Update Docker Repository & Build Docker Image

```
cd $HOME/Documents  
git clone https://github.com/TeMU-BSC/ictusnet-ctakes.git  
cd $HOME/Documents/ictusnet-ctakes  
# mkdir -p $HOME/Documents/ictusnet-ctakes/spactes-resources-bin/lib  
# mkdir -p $HOME/Documents/ictusnet-ctakes/spactes-resources-bin/config
```

**#IMPORT NOTE:**

**#We need just a few libraries, so:**



```

cd $HOME/Documents/ctakes/ctakes-SpaCTeS/target/lib

cp -rf commons-lang3-3.7.jar ctakes-core-4.0.1-SNAPSHOT.jar ctakes-core-res-4.0.1-SNAPSHOT.jar ctakes-fuzzy-dictionary-lookup-4.0.1-SNAPSHOT.jar ctakes-heideltime-4.0.1-SNAPSHOT.jar ctakes-openminted-freeling-4.0.1-SNAPSHOT.jar ctakes-SpaCTeS-res-4.0.1-SNAPSHOT.jar ctakes-spellcheck-core-1.1.2.jar ctakes-type-system-4.0.1-SNAPSHOT.jar ctakes-utils-4.0.1-SNAPSHOT.jar
de.tudarmstadt.ukp.dkpro.core.api.featurepath-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.api.io-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.api.lexmorph-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.api.metadata-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.api.parameter-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.api.resources-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.api.segmentation-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.api.syntax-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.io.text-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.io.xmi-asl-1.9.0.jar
de.tudarmstadt.ukp.dkpro.core.textcat-asl-1.9.0.jar ivy-2.4.0.jar jFreeing-4.1.jar log4j-1.2.17.jar lucene-core-4.0.0.jar plexus-utils-2.0.6.jar uimafit-core-2.4.0.jar uimafit-cpe-2.4.0.jar uimaj-adapter-vinci-2.10.2.jar uimaj-core-2.10.2.jar uimaj-cpe-2.10.2.jar uimaj-document-annotation-2.10.2.jar $HOME/Documents/ictusnet-ctakes/spactes-resources-bin/lib

cp -rf $HOME/Documents/ctakes/ctakes-SpaCTeS/target/config $HOME/Documents/ictusnet-ctakes/spactes-resources-bin

cp -rf $HOME/Documents/ctakes/ctakes-SpaCTeS/target/ctakes-SpaCTeS-4.0.1-SNAPSHOT.jar $HOME/Documents/ictusnet-ctakes/spactes-resources-bin

cd $HOME/Documents/ictusnet-ctakes

# Creating docker image

sudo bash build-docker.sh
    
```

## Push Image to Docker Hub

```

# NOTE: Just for first time, for the next time, every time we push in ictusnet repository
https://github.com/TeMU-BSC/ictusnet\_ctakes.git, Docker hub generate a new docker image.

# Create an account in https://hub.docker.com/

sudodocker login --usernameUSERNAME--password PASSWORD

# For Provide a password using STDIN
https://docs.docker.com/engine/reference/commandline/login/#provide-a-password-using-stdin

# Check the image ID using

sudodocker images

# and what you will see will be similar to

REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
ictusnet             latest              0bb11378728f       25 minutes ago     9.33GB
taln/freeling        latest              a3d8a4b9a104       2 years ago         6.71GB

# Add a tag in the image (latest)

sudodocker tag 0bb11378728f yourhubusername/ictusnet:latest
    
```

```
# Push the image to the repository you created
sudodocker push yourhubusername/ictusnet
# Explanation about how to use it (Add them into the docker hub -> readme):
# Pre-annotation dockerized pipeline for the ICTUSnet project
## Previous considerations
- `/full/path/to/your/input/data` is your input directory that
should be as below format:
` ``
data
├─ file1.txt
├─ file2.txt
└─ ...
` ``
- `/full/path/to/your/output/data` should be an empty directory to
put the `.ann` files that `ictusnet` docker container will
generate.
## Run
` ``
$ docker run --rm -it \
-v /full/path/to/your/input/data:/ICTUSnet/data/TXT \
-v /full/path/to/your/output/data:/ICTUSnet/data/ANN_FINAL \
bsctemu/ictusnet process.sh
` ``
```

### Run Docker Image

*# Note: Run the below script with two arguments: one path for TXT directory as an input and one path for ANN directory as an output.*

```
bash run-docker.sh /full/path/to/your/input/data
/full/path/to/your/output/data
```

### Build Singularity Image

*# Note: For building the Singularity image, once the Docker one has been built (since it is converted using <https://github.com/singularityhub/docker2singularity>), assuming Singularity is installed and Docker is a non-root user, run:*

```
bash build-singularity.sh
```

### Run Singularity Image

*# Note: Note: Run the below script with two arguments: one path for TXT directory as an input and one path for ANN directory as an output.*

```
bash run-singularity.sh /full/path/to/your/input/data
/full/path/to/your/output/data
```

### Pushing Changes to SpaCTeS repository

- `cd $HOME/Documents/ctakes/`
- `mvn clean -Dpg.skip -Dmaven.test.skip=true -DskipTests=false`
- `cp -rfctakes-core ctakes-openminted-freelingctakes-SpaCTeS-res ctakes-type-system pom.xml ctakes-fuzzy-dictionary-lookup ctakes-heideltimectakes-SpaCTeSctakes-spellcheck-service $HOME/Documents/spactes`
- `cd $HOME/Documents/spactes`
- `git add .`
- `git commit -m "changes ..."`
- `git push origin master`

## ANNEX 6 ICTUSnet cTAKES Developing Guidelines

### Introduction

We have implemented a clinical text-processing tool<sup>1</sup> for Spanish and Catalan EHRs (Electronic Health Records). This tool, the first in Spanish for neuroscientific purposes, was used to process a collection of anonymized discharge reports collected through a network that integrates 46 Catalan hospitals. The tool was primarily developed to assist human experts in the process of systematically identifying and extracting relevant information from discharge reports to evaluate the quality of hospital care for patients with a diagnosis of stroke. The tool is a pipeline that integrates three components in cTAKES (clinical Text Analysis and Knowledge Extraction System is an open-source Natural Language Processing system that extracts clinical information from electronic health record unstructured text):

1. FREELING (Padro and Stanilovsky, 2012) is a C++ library providing language analysis functionalities (Morphological Analysis, Named Entity Detection, PoS-Tagging, Parsing, Word Sense Disambiguation, Semantic Role Labelling, so forth) for a variety of languages. FREELING can be integrated into UIMA using a wrapper and a dockerized version of Freeling that was developed during the OpenMinTeD project.
2. HEIDELTIME (Strotgen and Gertz, 2010) is a multilingual, domain-sensitive temporal tagger that extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard.
3. FUZZY DICTIONARY LOOKUP identifies terms in the text and normalizes them to codes in a given ontology. This component is based on the Fast Dictionary lookup of cTAKES. Although the lookup algorithm has been changed completely. The Fast Dictionary Lookup component of cTAKES is strict to finding the matched words in the dictionary/lexicon, and therefore if in the input's EHR we have typos or missed tokens, the fast dictionary lookup component could not detect these tokens.

---

<sup>1</sup><http://github.com/TeMU-BSC/spactes/>

siabar Removing ctakes-fuzzy-dictionary-lookup-res		9c01ff5 11 days ago	🕒 43 commits
📁 ctakes-SpaCTeS-res	improving dic		11 days ago
📁 ctakes-SpaCTeS	Keeping the longest span for the same category		22 days ago
📁 ctakes-core	improving dic		11 days ago
📁 ctakes-fuzzy-dictionary-lookup	Removing ctakes-fuzzy-dictionary-lookup-res		11 days ago
📁 ctakes-heideltime	adding more variable on fuzzy dictionary		7 months ago
📁 ctakes-openminted-freeling	improving dic		11 days ago
📁 ctakes-spellcheck-service	Reading input and output from args		8 months ago
📁 ctakes-type-system	applying a function to pre-processing dictionaries		8 months ago
📄 .gitattributes	correcting project language		15 months ago
📄 LICENSE	increasing processing speed, reducing memory usage		9 months ago
📄 README.md	Adding BRAT Analysis Engine into cTAKES-Core		15 months ago
📄 pom.xml	Removing ctakes-fuzzy-dictionary-lookup-res		11 days ago

We integrated all of these components into cTAKES as native components

Note: Type System of cTAKES has been updated.

All of UIMA classes have two main methods, *Initialize* and *Process*:

### 1. Initialize

```
public void initialize(UimaContextContext)
```

```

public void initialize(UimaContext aContext) throws ResourceInitializationException {
    super.initialize(aContext);

    System.loadLibrary("Jfreeling");
    Util.initLocale("default");
    getLogger().info("Freeling, autodetect mode: " + autodetect + ", loading Spanish config");

    if (!this.autodetect) {
        try {
            init(language);
        } catch (Exception e) {
            e.printStackTrace();
            throw new ResourceInitializationException();
        }
    } else {
        lgid = new LangIdent(DATA + "common/lang_ident/ident.dat"); // ident-few for less languages!
    }
}

```

Which load all needed resource/parameters into the memory

## 2. Process

After running the initialize method of all components, based on the order of components in the pipeline, the *process* method can be executed.

```
public void process(JCas cas)

@Override
public void process(JCas cas) throws AnalysisEngineProcessException {
    String text = cas.getDocumentText();
    if (autodetect) {
        // language=cas.getDocumentLanguage();
        if (language == null || language.equalsIgnoreCase("x-undefined")) {
            // take the first 800 characters (200 words) Find what could be an end of
            // sentence

            // language = lgid.identifyLanguage(subStringForDetion(text));
            if (language.equals("none")) {
                getLogger().error("FreeLing, error in language detection, skip document");
                return;
            }
            getLogger().info("FreeLing, the language detected for document is: " + language);
            cas.setDocumentLanguage(language);

            cas.setDocumentLanguage(language);
        }
        try {
            init(language);
        } catch (Exception e) {
            getLogger().error("FreeLing, error initializing language, skip document");
            return;
        }
    }
    process(cas, text.substring(0, text.length()), 0);
}
```

### ctakes-SpaCTeS

The main class of this [file](#)<sup>2</sup> contains commands and parameters to run the ICTUSnet-ctakes (SpaCTeS) pipeline.

---

2

spactes/ctakes-

SpaCTeS/src/main/java/org/apache/ctakes/spactes/pipeline/SpaCTeSBuilderRunner.java

a

```
public static void main(final String... args) {
    try {
        LOGGER.info("StaCTeSBuilderRunner");

        String INPUT_DIR = args[0];
        String OUTPUT_DIR = args[1];

        PipelineBuilder builder = new PipelineBuilder();
        builder.readFiles(INPUT_DIR)

            // Segment Annotator from cTAKES
            .add(SimpleSegmentAnnotator.class)

            // FreeLing Wrapper (Tokenzier + POS + Lemma+ Sentence)
            .add(FreeLingWrapper.class, Collections.emptyList(),
                FreeLingWrapper.PARAM_LANGUAGE, "es",
                FreeLingWrapper.PARAM_DO_DEPENDENCY_PARSING, false,
                FreeLingWrapper.PARAM_USE_RULE_BASED, false,
                FreeLingWrapper.PARAM_LANGUAGE_AUTODETECT, false,
                FreeLingWrapper.PARAM_LEMMA, ConfigParameterConstants.LemmaForm)

            // Temporal Tagging from Heidelberg
            .add(Heidelberg.class, Collections.emptyList(),
                Heidelberg.PARAM_DEBUG, false,
                Heidelberg.PARAM_TEMPNYMS, false,
                Heidelberg.PARAM_LOCALE, "",
                Heidelberg.PARAM_DATE, true,
                Heidelberg.PARAM_DURATION, true,
                Heidelberg.PARAM_SET, false,
                Heidelberg.PARAM_TIME, true,
                Heidelberg.PARAM_GROUP, false,
                Heidelberg.PARAM_TYPE_TO_PROCESS, "narratives",
                Heidelberg.PARAM_LANGUAGE, "spanish")

            // Fuzzy Dictionary LookUp
            .addDescription(DefaultJCasTermAnnotator.createAnnotatorDescription(DICT_DESC_PATH,

        builder
        .writeXMIs(OUTPUT_DIR)
        // Brat Writer Component
        .writeBrat(OUTPUT_DIR);
```

Input is a text file and output is a XML(Readable by UIMA CVS), BRAT or HTML file.  
The input/output directories correspond to args[0] and args[1] arguments, respectively.

In this file, we can call all needed components by the .add method. Each component has several arguments, For example:

- **FreeLing:**  
PARAM\_LANGUAGE, "es",

PARAM\_DO\_DEPENDENCY\_PARSING, false,  
PARAM\_USE\_RULE\_BASED, false,  
PARAM\_LANGUAGE\_AUTODETECT, false.

- **HeidelTime**

PARAM\_DEBUG, false,  
PARAM\_TEMPONYMS, false,  
PARAM\_LOCALE, "",  
PARAM\_DATE, true,  
PARAM\_DURATION, true,  
PARAM\_SET, false,  
PARAM\_TIME, true,  
PARAM\_GROUP, false,  
PARAM\_TYPE\_TO\_PROCESS, "narratives",  
PARAM\_LANGUAGE, "spanish".

- **Fuzzy Dictionary Lookup**

DICT\_DESC\_PATH =  
"org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDictSpec.xml"  
In this xml file we indicate where is our Ictusnet dictionary:  
"org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDict.bsv"



```

<?xml version="1.0" encoding="UTF-8"?>

<lookupSpecification>
  <dictionaries>
    <dictionary>
      <name>LabAnnotatorTestDict</name>
      <implementationName>org.apache.ctakes.dictionary.lookup2.dictionary.BsvRareWordDictionary
      </implementationName>
      <properties>
        <property key="bsvPath" value="org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDict.bsv"/>
      </properties>
    </dictionary>
  </dictionaries>

  <conceptFactories>
    <conceptFactory>
      <name>LabAnnotatorTestConcepts</name>
      <implementationName>org.apache.ctakes.dictionary.lookup2.concept.BsvConceptFactory</implementationName>
      <properties>
        <property key="bsvPath" value="org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDict.bsv"/>
      </properties>
    </conceptFactory>
  </conceptFactories>

  <!-- Defines what terms and concepts will be used -->
  <dictionaryConceptPairs>
    <dictionaryConceptPair>
      <name>LabAnnotatorPair</name>
      <dictionaryName>LabAnnotatorTestDict</dictionaryName>
      <conceptFactoryName>LabAnnotatorTestConcepts</conceptFactoryName>
    </dictionaryConceptPair>
  </dictionaryConceptPairs>

  <rareWordConsumer>
    <name>Term Consumer</name>
    <implementationName>org.apache.ctakes.dictionary.lookup2.consumer.DefaultTermConsumer</implementationName>
    <properties>
      <property key="codingScheme" value="custom"/>
    </properties>
  </rareWordConsumer>

</lookupSpecification>

```

Fig 1: Dictionary Description XML file of Fuzzy Lookup Dictionary

IctusnetDict.bsv should have 5 columns and separated with a vertical bar: "|".

**SNOMED CT|ARCHETYPE|VARIANT|PREFERRED TERM|TYPO**

NOTE: VARIANTS should be unique and without accent. Accent from ctusnetDict.bsv can be removed by running main method RemoveAccents class<sup>3</sup>, with following args: Args[0] should be path of dictionary variable (IctusnetDict.bsv)

<sup>3</sup><https://github.com/TeMU-BSC/spactes/blob/master/ctakes-openminded-freeling/src/main/java/org/apache/ctakes/freeling/RemoveAccents.java>

Args[1] should be the following path in your local:

.../ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup/fuzzy

Both of these resources are saved in ctakes-SpaCTeS-res sub project.

### **ctakes-SpaCTeS-res**

In this sub project, we keep all necessary resources<sup>4</sup> for the components of the pipeline These include: IctusnetDictSpec.xml, IctusnetDict.bsv (which were explained in the previous section), and the needed resources for SpellChecker (lexicon.txt and dic.txt)

```
ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDict.bsv
ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDictSpec.xml
ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup/spellchecker/dic/dic.txt
ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup/spellchecker/lexicon/lexicon.txt
```

### **ctakes-openminded-freeling**

In this sub-project (component), we apply tokenizer and pos tagger. All components in the pipeline should use Freeling tokenizer, for this reason we defined a *tokenizer* method that can called from other components:

```
public List<String> tokenizer(String line) {
    // converting strange white space to normal space
    line = line.replace(" ", " ");
    ListWord l = tks.get(language).tokenize(line);
    // Split the tokens into distinct sentences.
    // getLogger().info(" sentence split" );
    ls = sps.get(language).split(sids.get(language), l, true);
    // Perform morphological analysis
    // getLogger().info(" morpho" );
    mfs.get(language).analyze(ls);
    // Perform part-of-speech tagging.
    // getLogger().info(" POS" );
    tgs.get(language).analyze(ls);
}
```

---

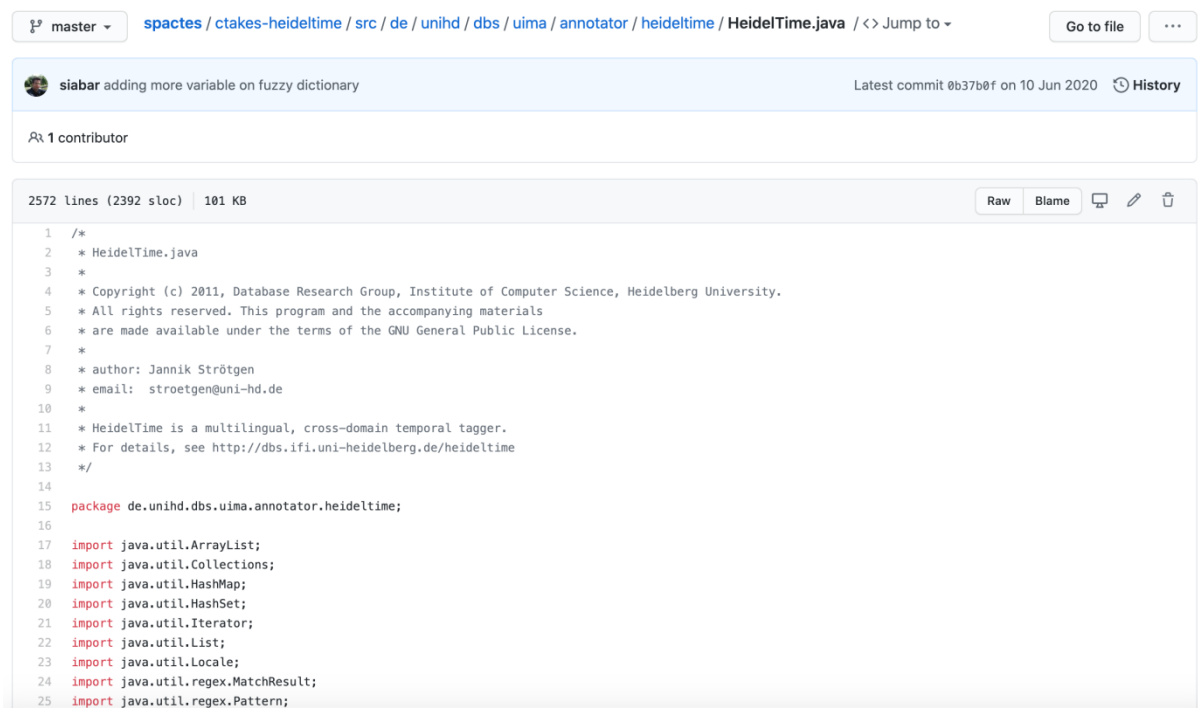
<sup>4</sup> <https://github.com/TeMU-BSC/spactes/tree/master/ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup>

All of the codes are in this [UIMA class](#)<sup>5</sup>.

### ctakes-heideltime

HeidelTime is a multilingual, domain-sensitive temporal tagger. For our project, we used EHT-TTS grammar which is HeidelTime grammar for temporal tagging of Spanish Electronic Health Records (EHR). The UIMA class of heideltime is in [this file](#)<sup>6</sup>.

All of the information will be saved in Timex3 typesystem (See further details in the ctakes-typesystem section).



```

1  /*
2  * HeidelbergTime.java
3  *
4  * Copyright (c) 2011, Database Research Group, Institute of Computer Science, Heidelberg University.
5  * All rights reserved. This program and the accompanying materials
6  * are made available under the terms of the GNU General Public License.
7  *
8  * author: Jannik Strötgen
9  * email: stroetgen@uni-hd.de
10 *
11 * HeidelbergTime is a multilingual, cross-domain temporal tagger.
12 * For details, see http://dbs.ifi.uni-heidelberg.de/heideltime
13 */
14
15 package de.unihd.dbs.uima.annotator.heideltime;
16
17 import java.util.ArrayList;
18 import java.util.Collections;
19 import java.util.HashMap;
20 import java.util.HashSet;
21 import java.util.Iterator;
22 import java.util.List;
23 import java.util.Locale;
24 import java.util.regex.MatchResult;
25 import java.util.regex.Pattern;
    
```

### ctakes-fuzzy-dictionary-lookup

The Fast Dictionary Lookup component of cTAKES (Native component in cTAKES) is very strict for finding the matched words in the dictionary/lexicon, therefore if in the input document we have typos or missed tokens, the dictionary lookup component cannot detect these tokens. For this reason we decided to completely modify the fast dictionary lookup component and integrate it with the SNOMED's SpellChecker. The SNOMED's SpellChecker uses Lucene similarity and it is very flexible for detecting typos, which are very usual in medical/clinical documents. We call the new component **Fuzzy Dictionary Lookup**.


<sup>5</sup> ctakes-openminted-freeling/src/main/java/org/apache/ctakes/freeling/FreeLingWrapper.java


<sup>6</sup> ctakes-heideltime/src/de/unihd/dbs/uima/annotator/heideltime/HeidelbergTime.java

First of all, the Fuzzy Dictionary Lookup tokenizes all words in dictionaries and lexicons based on Freeling's tokenizer and removes all accents. Also it converts words to lowercase if the given word's length is bigger than 1<sup>7</sup>. This restriction helps to distinguish between abbreviations and medical keywords such as "I" entities.

We used the Spellchecker of SNOMED that checks words against a given lexicon and dictionary<sup>8</sup> and gets suggestions for misspelt words. The algorithm goes as follows:

- First, the spellchecker would be apply with 70% accuracy just for the first word of the given text to lookup top similar medical word.
- If similar word has been found, the spell checker with 80% accuracy would be apply for all the given text and if it has been matched with one of the medical variants in the medical dictionary, the given text would be annotated with the correct archetype and SNOMED CT code.

 master ▾ [spactes](#) / [ctakes-fuzzy-dictionary-lookup](#) / [src](#) / [main](#) / [java](#) / [org](#) / [apache](#) / [ctakes](#) / [dictionary](#) / [lookup2](#) / Go to file Add file ▾

 **siabar** improving dic 36b64b3 21 hours ago [History](#)

..		
ae	improving dic	21 hours ago
concept	improving dic	21 hours ago
concurrent	First Version of SpaCTeS	15 months ago
consumer	First Version of SpaCTeS	15 months ago
dictionary	Improving fuzzy dictionary lookup	2 months ago
relation	First Version of SpaCTeS	15 months ago
term	First Version of SpaCTeS	15 months ago
textspan	First Version of SpaCTeS	15 months ago
util	improving dic	21 hours ago

The UIMA file of this component is [here](#)<sup>9</sup>. As mentioned before, the Fuzzy Dictionary Lookup uses the SNOMED's SpellChecker. The SpellChecker's resources (dictionary and lexicon) files are set in [this class](#)<sup>10</sup>.

<sup>7</sup> In the early versions it was 3, but for instance there are all possible cases of "mRS" such as mrs, MRS mrS and so on in the EHRs,. Therefore to detect all possible cases of this type of abbreviation if the word's length is 1 I kepted the original cases (such as D, I, E), otherwise it is converted to lowercase.

<sup>8</sup> This lexicon is all words in the medical dictionary

<sup>9</sup>ctakes-fuzzy-dictionary-lookup/src/main/java/org/apache/ctakes/dictionary/lookup2/ae/AbstractJCasTermAnnotator.java

<sup>10</sup>ctakes-fuzzy-dictionary-lookup/src/main/java/org/apache/ctakes/dictionary/lookup2/ae/DefaultJCasTermAnnotator.java

Note that for “ESCALAS<sup>11</sup>” variables, we did not use the SpellChler because of their nature, instead we applied a rule-based algorithm to detect them.

### **ctakes-spellcheck-service**

SNOMED's SpellChecker is a critical component for the project, especially because we deal with typos and need a relaxed matching mechanism against our lexicon (the one used in the Fuzzy Dictionary Lookup component). As it has been explained in ctakes-fuzzy-dictionary-lookup section, Spellchecker needs two files (dic.txt and lexicon.txt) for detect all possible typo, [lexicon.txt file](#) contains only first word of variants and [dict.txt](#) file contain all of variants in the [clinical dictionary](#)<sup>12</sup>. All of words in both of these files should be tokenized by running the main method of FreeLingWrapper class<sup>13</sup> with the following args:

args[0] should be input file<sup>14</sup>.

args[1] should be outfile file as output\_lexicon<sup>15</sup>

args[2] should be outfile file as output\_dic<sup>16</sup>

[SpellCheker](#) indexes a lexicon and dictionary in Lucene<sup>17</sup> and uses Lucene similarity to return top similarity over accuracy 0.83.

---

<sup>11</sup> “Escalas” variables are special case of variables in the ICTUSnet project. They include “ASPECTS (de 1 a 10): 10”, “NIHSS: 0-0-0-0- 0/0-0-0-0-0/0-0-0-1-0= 1”, “PUNTUACION TOTAL NIH:19”, “Escala NIHSS 0” or “mRS2”

<sup>12</sup> <https://github.com/TeMU-BSC/spactes/blob/master/ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDict.bsv>


<sup>13</sup> <https://github.com/TeMU-BSC/spactes/blob/master/ctakes-openminted-freeling/src/main/java/org/apache/ctakes/freeling/FreeLingWrapper.java>

<sup>14</sup> <https://github.com/TeMU-BSC/spactes/blob/master/ctakes-SpaCTeS-res/src/main/resources/org/apache/ctakes/examples/dictionary/lookup/fuzzy/IctusnetDict.bsv>



<sup>15</sup> <https://github.com/TeMU-BSC/spactes/blob/master/ctakes-SpaCTeS/org/apache/ctakes/examples/dictionary/lookup/spellchecker/lexicon/lexicon.txt>

<sup>16</sup> <https://github.com/TeMU-BSC/spactes/blob/master/ctakes-SpaCTeS/org/apache/ctakes/examples/dictionary/lookup/spellchecker/dic/dic.txt>

<sup>17</sup> <https://lucene.apache.org/>

 **siabar** increasing processing speed, reducing memory usage Latest commit fa3157e on Apr 27 [History](#)

1 contributor


120 lines (92 sloc) | 3.79 KB Raw Blame  

```



1 package org.ihtsdo.otf.spellcheck.service;
2
3 import org.apache.lucene.analysis.standard.StandardAnalyzer;
4 import org.apache.lucene.index.IndexWriterConfig;
5 import org.apache.lucene.search.spell.PlainTextDictionary;
6 import org.apache.lucene.search.spell.SpellChecker;
7 import org.apache.lucene.store.RAMDirectory;
8 import org.apache.log4j.Logger;
9
10 //import org.slf4j.Logger;
11 //import org.slf4j.LoggerFactory;
12
13 import java.io.File;
14 import java.io.FileReader;
15 import java.io.IOException;
16 import java.io.InputStream;
17 import java.io.InputStreamReader;
18 import java.io.Reader;
19 import java.util.*;
20 import java.util.regex.Matcher;
21 import java.util.regex.Pattern;
22
23 public class SpellCheckService {
24     ..
    
```

### ctakes-core

A class for writing annotations in a Brat file has been added to cTAKES in cTAKES-core project. Output of the pipeline can be saved in XML(Readable by UIMA CVS), HTML or a BRAT format. The native cTAKES only supported XML but [BRAT](#) format has been updated and also [HTML](#) format does not support Heidelberg CAS that It also has been updated as well.

 **siabar** improving dic Latest commit 36b64b3 21 hours ago [History](#)

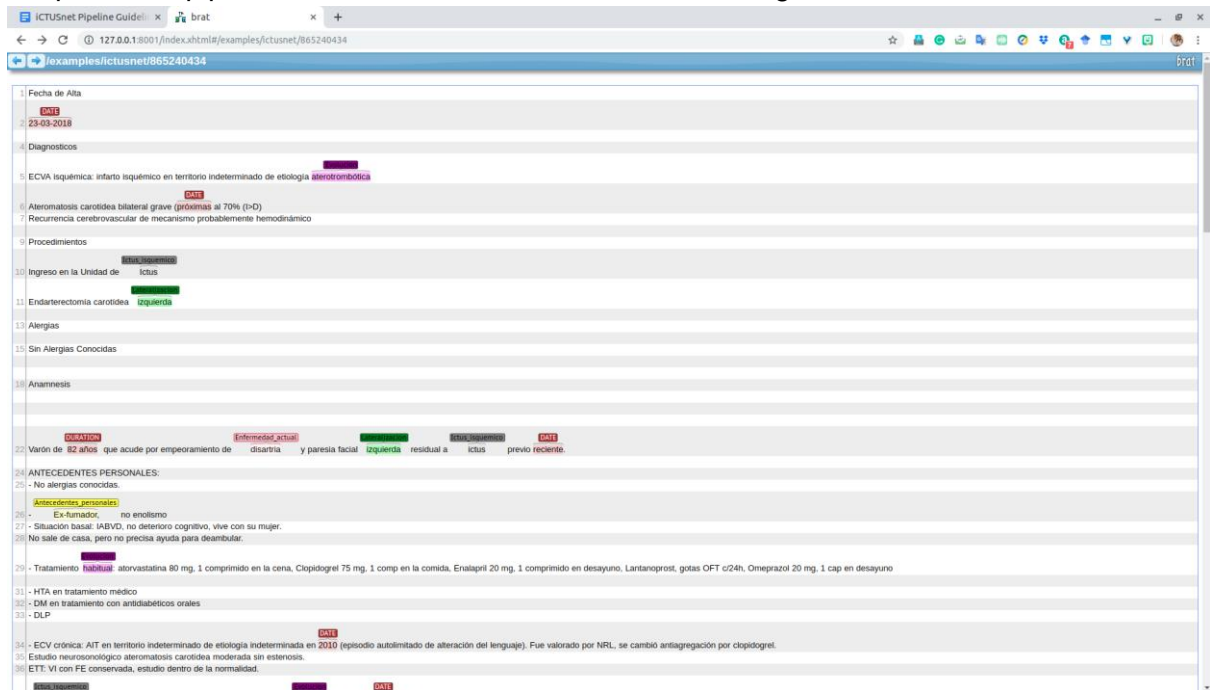
1 contributor

876 lines (791 sloc) | 32.6 KB Raw Blame  

```

1 package org.apache.ctakes.core.cc.brat;
2
3 import org.apache.ctakes.core.cc.AbstractJCasFileWriter;
4 import org.apache.ctakes.core.cc.pretty.SemanticGroup;
5 import org.apache.ctakes.core.pipeline.PipeBitInfo;
6 import org.apache.ctakes.core.util.DocumentIDAnnotationUtil;
7 import org.apache.ctakes.core.util.OntologyConceptUtil;
8 import org.apache.ctakes.core.util.textspan.DefaultTextSpan;
9 import org.apache.ctakes.core.util.textspan.OriginalTextSpan;
10 import org.apache.ctakes.core.util.textspan.TextSpan;
11 import org.apache.ctakes.typesystem.type.heideltime.Timeex3;
12 import org.apache.ctakes.typesystem.type.refsem.Event;
13 import org.apache.ctakes.typesystem.type.refsem.EventProperties;
14 import org.apache.ctakes.typesystem.type.refsem.UmlsConcept;
15 import org.apache.ctakes.typesystem.type.relation.BinaryTextRelation;
16 import org.apache.ctakes.typesystem.type.syntax.BaseToken;
17 import org.apache.ctakes.typesystem.type.textsem.*;
18 import org.apache.ctakes.typesystem.type.textspan.Segment;
19 import org.apache.ctakes.typesystem.type.textspan.Sentence;
20 import org.apache.log4j.Logger;
21 import org.apache.uima.fit.util.JCasUtil;
22 import org.apache.uima.jcas.JCas;
23 import org.apache.uima.jcas.tcas.Annotation;
24     ..
    
```

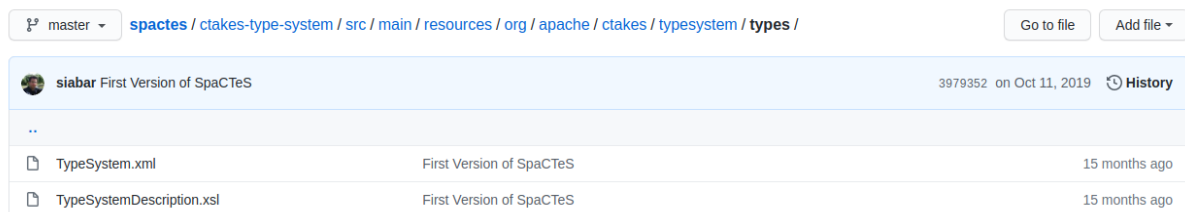
Output of the pipeline in BRAT format shows in the below figure:



## ctakes-type-system


For using non-native components into the pipeline, we need to add the new required types of these components into the [ctake-type-system](#) file and into the [type directory](#).

Note: For using CAS Visual Debugger (CVD), we need to load [typesystem file](#).




[master](#) / [spactes](#) / [ctakes-type-system](#) / [src](#) / [main](#) / [resources](#) / [org](#) / [apache](#) / [ctakes](#) / [typesystem](#) / [types](#) / **TypeSystem.xml**
Go to file

---

 **siabar** First Version of SpaCTeS Latest commit 3979352 on Oct 11, 2019 [History](#)

---

 **1** contributor

---

3140 lines (3131 sloc) | 147 KB
 Raw Blame

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <!--
3
4 Licensed to the Apache Software Foundation (ASF) under one
5 or more contributor license agreements. See the NOTICE file
6 distributed with this work for additional information
7 regarding copyright ownership. The ASF licenses this file
8 to you under the Apache License, Version 2.0 (the
9 "License"); you may not use this file except in compliance
10 with the License. You may obtain a copy of the License at
11
12 http://www.apache.org/licenses/LICENSE-2.0
13
14 Unless required by applicable law or agreed to in writing,
15 software distributed under the License is distributed on an
16 "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY
17 KIND, either express or implied. See the License for the
18 specific language governing permissions and limitations
19 under the License.
20
21 -->
22 <typeSystemDescription xmlns="http://uima.apache.org/resourceSpecifier">
23 <name>org.apache.ctakes.typesystem.types.TypeSystem</name>
24 <description>This is a Apache cTAKES Common Type System for clinical NLP. It includes general types necessary to store annotations and interface with clinical
25 </description>
26 <version>1.0</version>
27 <vendor>Apache cTAKES</vendor>
28 <types>
29 <typeDescription>
30 <name>org.apache.ctakes.typesystem.type.dependency.Dependency</name>
    
```

github.com/TeMU-BSC/spactes/blob/master/ctakes-type-system/src/main/resources/org/apache/ctakes/typesystem/types/TypeSystem.xml

```

303 </featureDescription>
304 </features>
305 </typeDescription>
306 <typeDescription>
307 <name>org.apache.ctakes.typesystem.type.heideltime.Timex3</name>
308 <description/>
309 <supertypeName>uima.tcas.Annotation</supertypeName>
310 <features>
311 <featureDescription>
312 <name>filename</name>
313 <description/>
314 <rangeTypeName>uima.cas.String</rangeTypeName>
315 </featureDescription>
316 <featureDescription>
317 <name>sentId</name>
318 <description/>
319 <rangeTypeName>uima.cas.Integer</rangeTypeName>
320 </featureDescription>
321 <featureDescription>
322 <name>firstTokId</name>
323 <description/>
324 <rangeTypeName>uima.cas.Integer</rangeTypeName>
325 </featureDescription>
    
```

## References

ICTUSnet: E2.4Development of supervised categorization models, topic modelling and extraction of clinical information via cognitive computing.



- [1] Savova, Guergana K., et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." *Journal of the American Medical Informatics Association* 17.5 (2010): 507-513.
- [2] Padró, Lluís, and Evgeny Stanilovsky. "Freeling 3.0: Towards wider multilinguality." *LREC2012*. 2012.
- [3] Strötgen, Jannik, and Michael Gertz. "Heideltime: High quality rule-based extraction and normalization of temporal expressions." *Proceedings of the 5th International Workshop on Semantic Evaluation*. 2010.