A Section Identification Tool: towards HL7 CDA/CCR Standardization in Spanish Discharge Summaries

Iakes Goenaga[a], Xabier Lahuerta[a], Aitziber Atutxa[b], Koldo Gojenola[b]

HiTZ Basque Center for Language Technology
http://www.hitz.eus
University of the Basque Country (UPV/EHU), Spain

[a]*Faculty of Computer Science, Pº Manuel Lardizabal, 1 — 20018 Donostia-San Sebastián*
[b]*School of Engineering, Paseo Rafael Moreno Pitxitxi, 3 — 48013 Bilbao*

Abstract

Background. Nowadays, with the digitalization of healthcare systems, huge amounts of clinical narratives are available. However, despite the wealth of information contained in them, interoperability and extraction of relevant information from documents remains a challenge.

Objective. This work presents an approach towards automatically standardizing Spanish Electronic Discharge Summaries (EDS) following the HL7 Clinical Document Architecture. We address the task of section annotation in EDSs written in Spanish, experimenting with three different approaches, with the aim of boosting interoperability across healthcare systems and hospitals.

Methods. The paper presents three different methods, ranging from a knowledge-based solution by means of manually constructed rules to super-

Email addresses: iakes.goenaga@ehu.eus (Iakes Goenaga), xlahuerta001@ikasle.ehu.eus (Xabier Lahuerta), aitziber.atucha@ehu.eus (Aitziber Atutxa), koldo.gojenola@ehu.eus (Koldo Gojenola)

vised Machine Learning approaches, using state of the art algorithms like the Perceptron and transfer learning-based Neural Networks.

Results. The paper presents a detailed evaluation of the three approaches on two different hospitals. Overall, the best system obtains a 93.03% F-score for section identification. It is worth mentioning that this result is not completely homogeneous over all section types and hospitals, showing that cross-hospital variability in certain sections is bigger than in others.

Conclusions. As a main result, this work proves the feasibility of accurate automatic detection and standardization of section blocks in clinical narratives, opening the way to interoperability and secondary use of clinical data.

Keywords: Section Identification, Interoperability, Electronic Discharge Summaries, HL7 Clinical Document Architecture

---

1. Introduction

The outstanding advancement of Machine Learning (ML) technologies (e.g., Deep Learning) enable us to more efficiently harness the large amounts of data collected through healthcare processes such as clinical narratives in electronic health records (EHR) as well as electronic discharge summaries (EDS). EHRs contain a lifetime record of the patient's complete medical history, diagnoses and treatment, medications, allergies and immunizations, as well as radiology images and laboratory results [1]. EDSs are an essential document to communicate patient journey and care planning regarding *an hospitalization episode* to the next practitioner [2][1]. In 2016 the proportion of primary care practices using electronic clinical records was about 80% on

[1]Some authors use these two terms interchangeably.

average across 15 EU countries [3], and in 2020 in the US the percentage is of 96% [4]. Digitalization of healthcare systems is contributing to the improvement of clinical and translational studies, and interoperability and information exchange between healthcare systems is more necessary than ever. For that reason, public policies and recommendations are pushing onto that way [5, 6, 7].

There is an increasing interest for integrating heterogeneous health information for different reasons: to facilitate the cross-border interoperability of information among healthcare systems, federal states and countries to ensure that citizens can securely access and exchange their health data wherever they are, and also to make digital health information more usable to the bedside and beyond [5, 6]. Several standards as openEHR [8], HL7-FHIR [9], HL7 CDA/CCR [10] are examples of this standardization effort.

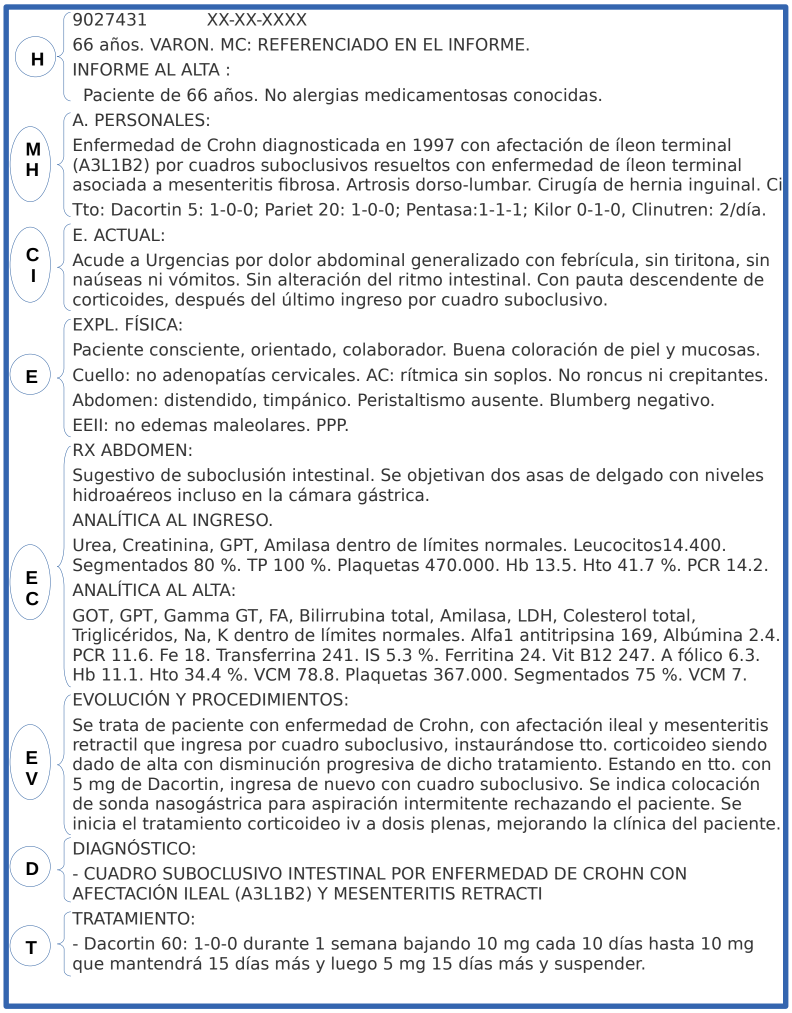
However, despite the wealth of information contained in the clinical narratives, interoperability and extraction of relevant information from documents remains a challenge. Although the aforementioned standards exist, so far they have not been widely adopted, and even if so, the healthcare system at large still has a huge amount of untapped legacy clinical text.

Healthcare systems provide guidelines for writing clinical documents, which for operative reasons typically follow some minimal principles to ensure the optimal interactions between health professionals and patients like SOAP (Subjective, Objective, Assessment, Plan), or APIE (Assessment, Plan, Implementation, and Evaluation) [11, 12]. Some systems assume that these principles are best reflected by using free text, due to flexibility to express anything that the health-care providers need to record. On the opposite extreme, some impose structured or semi-structured clinical documents in sections, where each section is a main block of information. In all cases, the automated processing of clinical texts is hampered by ambiguity, lexical variety, use of abbreviations, errors due to mistakes, redundancies, etc.

Under this scenario, this work presents a first approach towards automatically standardizing Spanish EDSs following the HL7 Clinical Document Architecture (CDA) R2 template for Discharge Summaries [10] for both helping interoperability and secondary use of Electronic Discharge Summaries.

The HL7 CDA R2 template contains a set of clinically relevant sections, and part of this standardization task is known as Section Identification. It is defined in [13] as *detecting the boundaries of text sections and adding semantic annotations*. They define a section as *a text segment that groups together consecutive clauses, phrases or sentences that share the description of one dimension of a patient, patient's interaction or clinical findings*. A section can be marked explicitly, through structural demarcations (headings or subheadings), or it can exist implicitly. The main assumption for making this identification is that unstructured texts have an explicit or implicit structure.

Besides its relevance in terms of standarization and interoperability, section identification provides a deeper understanding of EDSs, for instance, by recognizing the section in which a medical entity is located. The same medical condition found in the "past personal medical history" or in the "family medical history" section might lead to different conclusions. Several works on secondary use of EHRs and EDSs have shown that section identification can be helpful for a variety of tasks [14] such as entity recognition [15], cohort retrieval [16] and temporal relation extraction [17], and can help in most automatic medical processing tasks, as ICD-10 coding [18, 19, 20, 21]. This issue is rapidly becoming an important topic in both academia and industry.



H
9027431 XX-XX-XXXX
66 años. VARON. MC: REFERENCIADO EN EL INFORME.
INFORME AL ALTA :
Paciente de 66 años. No alergias medicamentosas conocidas.

MH
A. PERSONALES:
Enfermedad de Crohn diagnosticada en 1997 con afectación de íleon terminal (A3L1B2) por cuadros suboclusivos resueltos con enfermedad de íleon terminal asociada a mesenteritis fibrosa. Artrosis dorso-lumbar. Cirugía de hernia inguinal. Ci
Tto: Dacortin 5: 1-0-0; Pariet 20: 1-0-0; Pentasa:1-1-1; Kilor 0-1-0, Clinutren: 2/día.

CI
E. ACTUAL:
Acude a Urgencias por dolor abdominal generalizado con febrícula, sin tiritona, sin naúseas ni vómitos. Sin alteración del ritmo intestinal. Con pauta descendente de corticoides, después del último ingreso por cuadro suboclusivo.

E
EXPL. FÍSICA:
Paciente consciente, orientado, colaborador. Buena coloración de piel y mucosas.
Cuello: no adenopatías cervicales. AC: rítmica sin soplos. No roncus ni crepitantes.
Abdomen: distendido, timpánico. Peristaltismo ausente. Blumberg negativo.
EEII: no edemas maleolares. PPP.

EC
RX ABDOMEN:
Sugestivo de suboclusión intestinal. Se objetivan dos asas de delgado con niveles hidroaéreos incluso en la cámara gástrica.
ANALÍTICA AL INGRESO.
Urea, Creatinina, GPT, Amilasa dentro de límites normales. Leucocitos14.400. Segmentados 80 %. TP 100 %. Plaquetas 470.000. Hb 13.5. Hto 41.7 %. PCR 14.2.
ANALÍTICA AL ALTA:
GOT, GPT, Gamma GT, FA, Bilirrubina total, Amilasa, LDH, Colesterol total, Triglicéridos, Na, K dentro de límites normales. Alfa1 antitripsina 169, Albúmina 2.4. PCR 11.6. Fe 18. Transferrina 241. IS 5.3 %. Ferritina 24. Vit B12 247. A fólico 6.3. Hb 11.1. Hto 34.4 %. VCM 78.8. Plaquetas 367.000. Segmentados 75 %. VCM 7.

EV
EVOLUCIÓN Y PROCEDIMIENTOS:
Se trata de paciente con enfermedad de Crohn, con afectación ileal y mesenteritis retractil que ingresa por cuadro suboclusivo, instaurándose tto. corticoideo siendo dado de alta con disminución progresiva de dicho tratamiento. Estando en tto. con 5 mg de Dacortin, ingresa de nuevo con cuadro suboclusivo. Se indica colocación de sonda nasogástrica para aspiración intermitente rechazando el paciente. Se inicia el tratamiento corticoideo iv a dosis plenas, mejorando la clínica del paciente.

D
DIAGNÓSTICO:
- CUADRO SUBOCLUSIVO INTESTINAL POR ENFERMEDAD DE CROHN CON AFECTACIÓN ILEAL (A3L1B2) Y MESENTERITIS RETRACTI

T
TRATAMIENTO:
- Dacortin 60: 1-0-0 durante 1 semana bajando 10 mg cada 10 días hasta 10 mg que mantendrá 15 días más y luego 5 mg 15 días más y suspender.

Figure 1: Example EDS and its sections (H: Heading, MH: Medical History, CI: Current Illness, E: Exploration, EC: Complementary Exploration, EV: Evolution, D: Diagnosis, T: Treatment).

Given the difficulty in accurately extracting data from text, most non-research use of EHR and EDS data rely only on structured data. However, clinical notes contain highly valuable information not found in strictly structured fields and, moreover, they give access to volumes of data that are orders of magnitude bigger and, consequently, improving retrieval accuracy from text would have great value.

In this paper, we will explore the task of section annotation in EDSs written in Spanish (see Figure 1). We will experiment with three different approaches, ranging from a knowledge-based solution by means of manually constructed rules to supervised Machine Learning approaches, including the structured Perceptron algorithm and Deep Neural Networks. The paper will present a detailed evaluation of the three approaches and, as a main result, will prove the feasibility of automatically detecting section blocks in EDSs.

The main contributions of this work are:

- We describe an annotation format for EDSs that defines the section structure of a document. We have evaluated its feasibility annotating a dataset comprised of 300 documents and have measured a high inter-annotator agreement.
- We implement three different approaches to automatic section identification, including a rule-based method, the Perceptron online learning algorithm and Neural Networks.
- We conduct exhaustive experiments to explore the contribution of each method, also giving a detailed analysis of the strengths and weaknesses of the proposed approaches.

The remainder of this paper is structured as follows. Section 2 discusses related work. The resources and corpus are presented in Section 3. Section 4 sketches the main results, while Section 5 provides an analysis of the results including a comparison of the different approaches as well as an estimation of the system's ability to generalize across hospital settings and a qualitative evaluation of the encountered errors. Finally, Section 6 summarizes the main conclusions and future work.

2. Background

Pomares-Quimbaya et al. [13] reviewed several studies on clinical section identification, which varied on the kind of narrative, the type of section, and the application. The paper examines the characteristics of systems using a strategy for section identification, the methods used to identify implicit or explicit sections with different degrees of success, and the main application scenarios and contexts that have been used with good performance. From the technical point of view, the methods were classified into rule-based methods (59%), machine learning methods (22%) and a combination of both (19%). According to the authors, hybrid methods showed the best performance. 46% of the studies were able to identify explicit (using headings) and implicit sections. Regarding the language of application, most of the works (78%) were intended for English texts.

Arnold et al. [22] present SECTOR, a model to segment documents into sections, under the hypothesis that topics, learned in an unsupervised way, characterize semantically coherent text segments (sections). Their deep neural network architecture learns a latent topic embedding over a document, in order to classify local topics and to segment a document at topic shifts. They report a 56.7% F-score for segmentation and classification in the domain of diseases. Although the approach seems promising, its main inconvenient for our task is that, as topics are learned in an unsupervised manner, the topic clusters do not fit well with the nine HL7 section types of our documents, because topic clusters can be either finer or more coarse-grained.

Choi et al. [23] claim that the structure underlying EHR data improves the performance of prediction tasks such as heart failure prediction. As most EHR data do not always contain complete structure information or is completely unavailable, they experiment alternatives to the baseline consisting of treating EHR data as a flat-structured bag-of-features. The proposed model outperformed the baseline approach for various prediction tasks such as readmission and mortality prediction, indicating that the detection of EHR structure is beneficial for many tasks.

Rosenthal et al. [24] developed a system to detect sections in EHRs, based on different architectures: an RNN based system and a transfer based system using BERT. To overcome the lack of annotated data they propose to use for training purposes sections learned from medical literature (journals, textbooks, web content). They conclude that out of domain clinical literature is helpful when there is not enough EHR data, but its contribution is not significant with bigger sizes of the in-domain annotated dataset. Their system did not exploit the structure of the document, that is, the fact that sometimes sections are ordered in a canonical order (i.e., first the *Chief complaint*, then the *Antecedents*, ...), which we plan to use in our approach, as it can be helpful in deciding section types.

Rush et al. [25] solve the section identification problem using a CRF classifier to mark each token as belonging to a section header, and then they apply a rule-based post-processing module to structure the annotated sections. Comparing to our work, they do not perform normalization and therefore the number of sections they identify is not fixed. In their system, similar section headers are considered different, while our aim is to normalize each section into a set of nine HL7 section types.

Apart from the medical domain, other areas like legal decision-support systems also leverage the content structure of documents. For example, Branting et al. [26] exploit structural and semantic regularities in law case corpora to identify textual patterns that have both predictable relationships to case decisions and explanatory value for legal decision support and explainable outcome prediction.

To summarize, we can see that the identification of sections is currently a promising area of active research, specially for languages other than English. Historically, rule-based methods have been the most widely used approach, although the recent emergence of new ML and Deep Learning techniques that have revolutionized the state of the art on many tasks also presents avenues for new developments.

3. Materials and Methods

In this section we will explore all the corpora and tools we have used in order to carry out the experiments. In the first part (section 3.1), we present the annotated corpus, the defined annotation model and the inter-annotator agreement. Section 3.2 gives a description of the large unannotated corpora used as an additional resource to derive a language model for the Deep Learning approach (see subsection 3.3.3). Then, section 3.3 describes the three approaches that have been followed for the automatic annotation of sections in EDSs.

3.1. Annotated Corpus

3.1.1. The EDS Corpus

We chose to analyze clinical reports of long-term hospital discharges from two hospitals of the Osakidetza Health System, the Galdakao-Usansolo and Basurto hospitals. Discharge documents are issued by the responsible doctor in a health center at the end of each patient's healthcare process, specifying the patient's data, a summary of their clinical history, the healthcare activity provided, diagnosis and therapeutic recommendations. They are documents of great importance within the clinical history, containing the summary of the care provided to the patient during the hospitalization episode. The recipients are different users with diverse interests, including the patient, his/her family, the primary care physician and the specialist physician.

A set of 300 documents was selected for manual annotation (see Table 3) divided evenly between the two hospitals. As each EDS typically can contain most of the nine section types that we have defined, this corpus, albeit of a moderate size, can give a sufficient amount of data (more than 2,000 instances of the different section types) for training and evaluation.

3.1.2. Definition of the Main Section Types

As mentioned in the introduction, we have followed the HL7 CDA R2 recommendations proposed for EDSs. This standard requires EDSs to be minimally structured in at least three main sections: Hospital Course Section, Discharge Diagnosis Section and Plan of Treatment Section. Nevertheless, there are 22 other sections that are optional and try to cover the heterogeneity of information captured in EDSs. This optionality allows each healthcare system to select those sections that better accommodate to the reported information. In our case and following [27] we selected 9 sections out of those 22. Table 1 summarizes the adopted HL7 CDA R2 sections specially recommended for Electronic Discharge Summaries along with a short description and the nomenclature we will employ all over the paper.

Although ideally all EDSs could contain each and all of the listed sections, in practice this just does not happen most of the times. It is usual to find summaries with less elements and it can also happen to find some sections more than once in a document, possibly when the discharge summary includes more than one episode from one patient.

Regarding the order, although the given description of section types can be considered as a canonical ordering of the elements in an EDS, there is a great variability. Except for the heading, that appears almost always in the first position, the rest of the sections can be found in different parts of the document. For example, even if it is common to find the diagnosis and treatment at the end of the EDS, some practitioners tend to move them towards the beginning of the document. Even when many sections are marked by an explicit heading, there are several challenges related to the detection of section boundaries:

- Variability of section headings. Although the standard definition could suggest that all the headings are naturally defined by a common term,

HL7 CDA R2 name	**Abbreviated name**	**Description**
-	Header (H)	Not a HL7 CDA R2 section but included to capture the header, which can be highly specific to each hospital
Chief complaint	Chief complaint (CC)	This section is similar a press headline, and it briefly contains the answers to who, what, where, why and when
Past Medical History	Medical History (MH)	Past symptoms, medications, diseases or procedures. Sometimes there are specific subsections for *Family history* or *Personal history*.
History of Present Illness	Current Illness (CI)	A detailed description of the issues presented in the chief complaint section.
Vital Sign Section	Exploration (E)	This section describes observations, including the vital signs, muscle power and examination of different organs, especially ones that might be related to the symptoms.
Hospital Discharge Studies Summary	Compl. Exploration (EC)	Additional, specialized tests, like ECG, or a radiography.
Review of Systems	Evolution (EV)	Evolution of the patient during the hospitalization.
Discharge Diagnosis	Diagnoses (DI)	Main and secondary diseases diagnosed by a medical practitioner.
Plan of Treatment	Treatment (T)	Medications, procedures and recommendations for this patient

Table 1: Brief description of the HL7 CDA R2 adopted sections.

there is a great variability, corresponding to the use of synonyms, abbreviations and variations. Additionally, in some cases section headings can be misleading or ambiguous, and the content of the text accompanying the heading must be taken into account in order to disambiguate the text.

- Implicit sections. Apart from the headings, sections can also be identified by looking to the body text. For example, the description of measures like Sodium or Potassium will typically appear at the Exploration section. This information is also useful to detect sections, and is the only possibility when the section heading is not present. Table 2 shows how a considerable proportion of sections do not have an explicit section header. For example, only a 30% and 41% of the *Chief Complaint* and *Complementary exploration* sections, respectively, are explicitly marked. Some others, although they have an explicit heading most of the times, show a great variability (e.g., *EXPLORACION* in Table 2).


```xml
<?xml version = "1.0"?>
<sections>
<section id="1" type="ENCABEZADO" str="" offset="1-6"></section>
<section id="2" type="ANTECEDENTES" str="A. PERSONALES" offset="7-10"></section>
<section id="3" type="ENFERMEDAD ACTUAL" str="E. ACTUAL" offset="11-13"></section>
<section id="4" type="EXPLORACION" str="EXPL. FÍSICA" offset="14-19"></section>
<section id="5" type="EXPLORACION COMPLEMENTARIA" str="RX ABDOMEN" offset="20-29"></section>
<section id="6" type="DIAGNOSTICO" str="IMPRESIÓN DIAGNÓSTICA" offset="30-32"></section>
<section id="7" type="EVOLUCION" str="EVOLUCIÓN Y PROCEDIMIENTOS" offset="33-36"></section>
<section id="8" type="DIAGNOSTICO" str="DIAGNÓSTICO" offset="37-39"></section>
<section id="9" type="TRATAMIENTO" str="TRATAMIENTO" offset="40-50"></section>
</sections>
```


Figure 2: Example of EDS annotation corresponding to Figure 1.

We chose to use a stand-off annotation based on XML. For example, Figure 2 presents the annotation document for the EDS presented in Figure 1. Each section is described by an XML element containing attributes for the section type, the string (if any) that indicates the start of the section

Section	Examples
ENCABEZADO **(HEADING)**	9027431 16-04-09 66 años.
	VARON. MC: REFERENCIADO EN EL INFORME.
MOTIVO DE CONSULTA **(CHIEF COMPLAINT)**	MOTIVO DE INGRESO ...
	MOTIVO DE CONSULTA ...
	Paciente que ingresa procedente de ...
	Paciente varón de 47 años que ingresa para ...
	Varón de 87 años ingresado desde ...
	Varon de 63 años que consulta por ...
	MI ...
ANTECEDENTES **(MEDICAL HISTORY)**	ANTECEDENTES PERSONALES ...
	A.PERSONALES ...
	AP ...
	A. Personales ...
	A.P. ...
	A. PERSONALES ...
	Paciente de 65 años de edad con antecedentes ...
EXPLORACION **(EXPLORATION | PHYSICAL EXAMINATION)**	EXPLORACION GENERAL ...
	EXPLORACION ...
	Exploración física ...
	EXPLORACION VASCULAR ...
	EXPLORACIÓN ORL ...
	EXPLORACIÓN PSICOPATOLÓGICA ...
	EXPLORACIÓN CLÍNICA EN LA UNIDAD ...
EXPLORACION COMPLEMENTARIA **(COMPLEMENTARY EXPLORATION)**	PRUEBAS COMPLEMENTARIAS ...
	EXPLORACION COMPLEMENTARIA ...
	ECG ...
	RX ABDOMEN ...
	ANALITICA AL ALTA ...
	ANALÍTICA EN URGENCIAS ...

Table 2: Examples of some instances of the beginning of identified Section Types.

and the offset (in lines) corresponding to the text of the section[2].

[2]Although the string attribute was useful for annotators when discussing any annota-

After the main section types to be annotated were defined, a set of documents from two different hospitals was annotated by two annotators. Table 3 describes the three-way data split (training, development and test). In order to minimize the annotation effort, the number of annotated documents could not be too large but it should also provide enough data for training and evaluation. Taking these considerations into account, a corpus of 300 documents was randomly selected. Regarding the split of the dataset into three subsets corresponding to training, development (or validation), and the final test, the validation and test sets should contain enough instances of each section type for the evaluation to be significant. For this reason we decided that each of the three subsets would contain 100 documents, different from classical data splits (e.g., 70%, 15% and 15% for training, development and test, respectively). Figure 3 shows the distribution of sections in both the train and development splits. Note that all sections are represented in both splits and their distribution is similar.

tion disagreement, in fact the attributes that should be obtained by an automatic system will be the section type and its location (line offset) in the document.

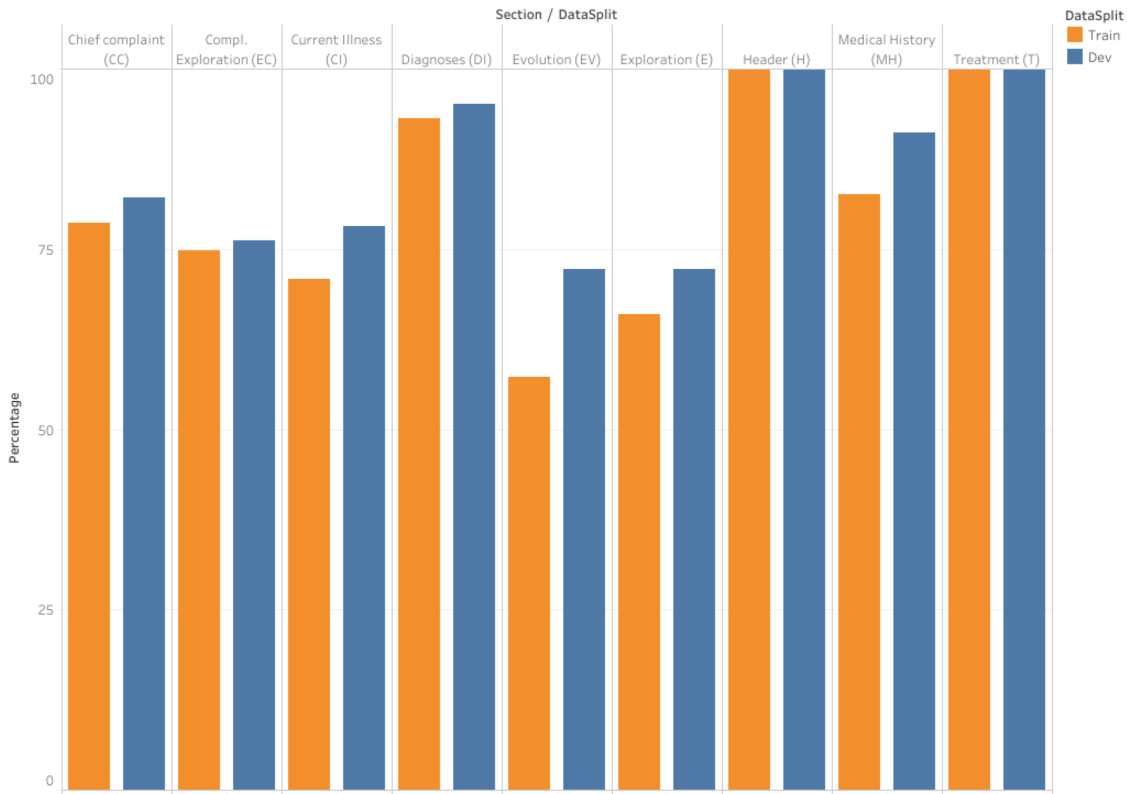


Figure 3: Distribution of sections in both train and development splits. Section chief complaint, for instance, is present in 82% of the development split, while slightly less in the training split (79%).

	Documents	**Sections**	**Tokens**
training	100	744	47,449
development	100	786	48,461
test	100	754	59,119

Table 3: Details of the annotated corpus.

3.1.3. Inter-Annotator Agreement

The annotation process of the corpus was performed by annotation experts. Annotators followed an iterative process of training until a high inter-annotator agreement was reached. The final agreement measure was calculated on a set of 25 EDSs that were doubly annotated by two different annotators, reaching a pairwise agreement of 93.47% Cohen's Kappa, indicating that the agreement is very high. There were differences with respect to each section type, ranging from 86% for Diagnosis (lowest agreement) to 100% for some section types, thus reaching a significant agreement for all section types.

Our annotation strategy requires each section type to be matched exactly while taking into account its content, and additionally returns the first and last lines of each section. While this strategy might seem an overly stringent criteria, the task is well defined as evidenced by the high inter-annotator agreement.

3.2. Textual Corpora

Deep learning techniques usually require huge amounts of data. Although manually annotated data gives the best results, it is very expensive and time consuming. For that reason, the idea of acquiring useful information in an unsupervised manner is very attractive, and efficient and effective methods have been developed. Vectorial representations of words, also known as *word embeddings* [28, 29], that are learned from textual corpora, have proven useful as an information source for many Natural Language Processing tasks, such as Part-Of-Speech (POS) tagging, Named Entity Recognition or Machine Translation, due to their ability to acquire relevant generalizations. These embeddings are learned through solving an appropriate optimization objective [28] under the assumption that similar words occur in similar contexts. As a result, vectors of similar words derived from such optimization tend to reside in the neighborhood in the vector space. There are two kinds of embeddings, static and contextual. Static embeddings capture in a vectorial representation information of a word form, while contextual embeddings are sensitive to context, representing both a word and its context.

This way, an unsupervised system can utilize the information based on word similarity in a manner that associates unseen words with those already occurring in the annotated corpus, thereby allowing us to cover unseen and misspelled terms. For instance, *infarct* and *stroke* are similar terms but one of them may not be in the annotated data set. The resulting word vectors will be fed to the neural network as input during training (see Figure 4), thus providing a model of the language that can help obtain better generalizations and, consequently, increase the recall of the final tool.

For this work we have employed heterogeneous embedding information both static and contextual in order to make the system sensitive to different granularity and domain specificity. Regarding the granularity, during the section identification training, adding a character embedding layer allows the system to learn at the character level. Besides the character embeddings learned during the training, we incorporated pre-trained, character-based embeddings based on fastText [30] trained over the Spanish version of Wikipedia. Character-based embeddings are able to generalize over n-grams, enabling the system to take into account prefixes and suffixes as well as to capture information about the different n-gram variations on the section heading words. They also generalize over zero-shot words, words that do not appear in the training corpus as their building element are characters and not words.

Table 4 presents the details of the different word embeddings we have used for the task. Static embeddings were obtained applying word2vec [30] to Electronic Discharge Summaries (50M words), together with pretrained embeddings that had been calculated with Wikipedia2Vec [31], representative of general domain. Additionally, we also used contextual string embeddings [32] we calculated from Electronic Discharge Summaries and Wikipedia .

Technique	**Source text**	**Embedding type**	**Details**
word2vec	EDS	static	window length = 1, dimensions = 300, algorithm = SkipNgram
Wikipedia2Vec	general domain		window length = 5, dimensions = 300, algorithm = Skipgram
FLAIR	EDSs	contextual	layers=1, hidden size = 2,048, sequence length = 250, mini batch size = 32
	general domain		layers = 1, hidden size = 1,024, sequence length = 250, mini batch size = 100

Table 4: Overview of the different embedding types used in this work (static word embeddings and contextual character embeddings).

3.3. Approaches to Automatic Section Identification

In this section we will explain the different approaches we have tried with the aim of automatically identifying sections in medical records. First of all, in subsection 3.3.1 we will specify the setup we used for the rule-based tool that we have developed. After that, in subsections 3.3.2, and 3.3.3 we will explore the ML algorithms we have employed, the Perceptron and Deep Learning, respectively. For both ML approaches, we have approached the task as a sequential learning process [33, 34], where the text is considered a sequence of tokens, and each token is associated with one tag indicating its corresponding section. We have used an IOB (Inside, Outside, Begin) tag model, where the beginning of each section is marked with a B tag (e.g., *B-DIA* for the token starting a diagnosis), the tokens inside a section are marked with an I tag (*I-DIA* will mark a token inside a diagnosis section), and using the O tag for elements that do not belong to any section (see Figure 5). This way, section identification can be viewed as the detection of extended and long entities. This approach has been successfully used in similar tasks as the identification of elementary discourse units (text segments consisting of one or several sentences) in Discourse processing [35] or topic segmentation [22]. Figure 4 presents an architecture of the system.

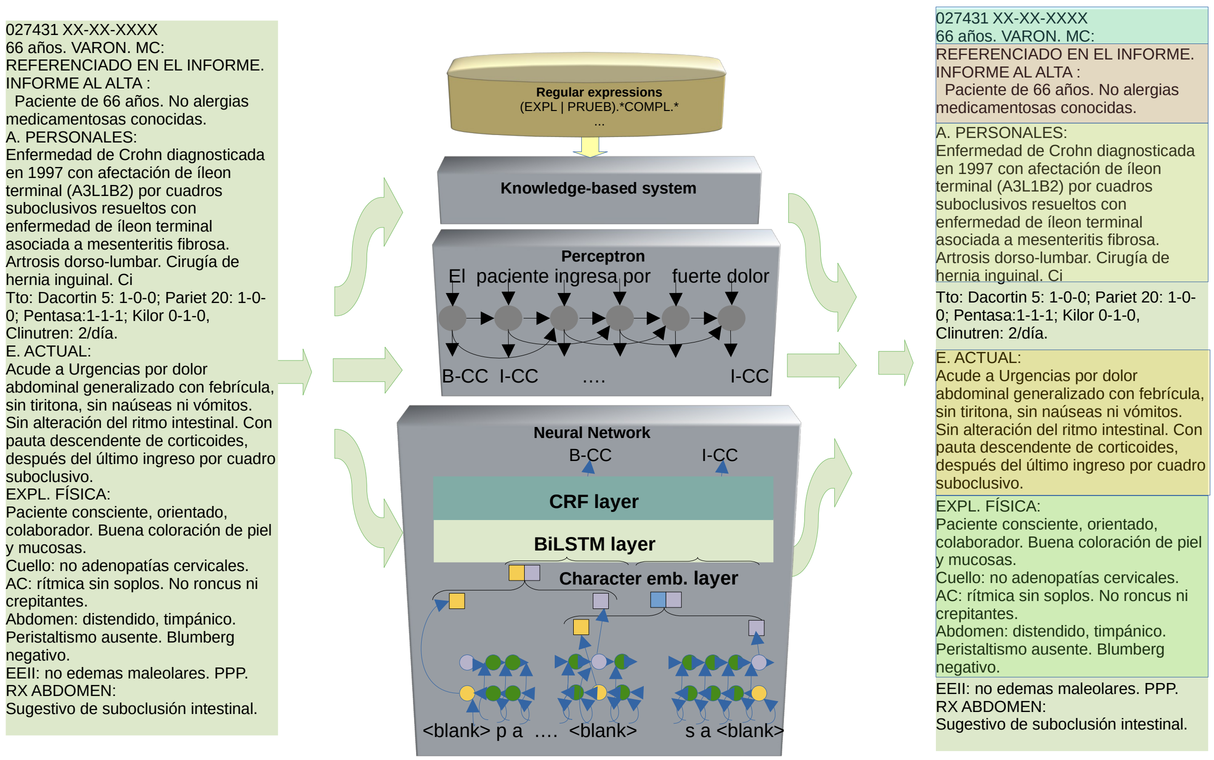


Figure 4: Architecture of the system. Three different approaches have been used: regular expressions, the Perceptron algorithm and neural networks.

3.3.1. Rule-Based Approach

Manually defined rules have been used since the early years of Artificial Intelligence, and are still a competitive method to achieve acceptable results. Their downside is the effort needed to include knowledge into the automatic system. Another drawback is their lack of generalization, because a change in the domain may imply a complete re-implementation of the rule system.

Regarding the identification of sections in medical records, this approach has been used in many systems [13, 36], where acceptable results have been reported, although in several cases the approach has not been general, but rather limited to a reduced set of very specific sections or portions of text. Table 2 presents several examples of the beginning of different section types. The table shows how there is a high variability difficult to capture using rules, specially with implicit sections with no standard title, like in the *Chief Complaint* and *Complementary Exploration*. Examples (1) and (2) present two rules that try to capture the start of the *Chief Complaint* and the *Current Illness* sections, where the parentheses enclose optional elements. The objective was to cover the different options found in the training set.

(1) MOTIVO(S) (DE(L)(A)) $INGRESO|PETICION|$
$EXPLORACION|ESTUDIO|CONSULTA\ (ACTUAL)$

(2) $(E.|ENFERMEDAD|SITUACIÓN|EPISODIO|ESTADO)$
$(A.|ACTUAL)\ |\ SINTOMATOLOGÍA$

3.3.2. Machine Learning: Perceptron

For the application of ML to section identification, we modeled the problem as a sequence to sequence problem. The task consists in learning to map from input word sequences $w_1 ... w_m \mid w_i \in W$ to output tag sequences $t_1 ... t_m \mid t_i \in T$.

Although some approaches to section identification used sentence sequences as input units [13], we preferred to model this problem using word sequences as input units to capture the fact that individual words in the right context are good signals for sections and also to reduce sparsity, because sentence sequences are more sparse than word sequences. The problem is cast as the assignment of the correct tag to each token. Although the tag assignment is made token by token, the final evaluation will be done on the detection of complete sections.

To do so, we employed the Averaged Structured Perceptron algorithm [37, 38] which combines the Perceptron algorithm for learning linear classifiers with an inference algorithm and converts a classification problem into a ranking problem. The objective of the algorithm is to find, for each sentence, the sequence of tags with the maximum score. This prediction decision process is divided into a sequence of smaller decisions made from left-to-right. Thus, at each step there is a word and its context, called *the history*, in which the local tagging decision is made, namely to predict the tag given the history. The history can be represented in several ways, using the prefixes of a given number of previous words, and/or the suffixes, or any other features that could be relevant for the task and then converted into a feature vector where each feature will get a weight through the learning process.

Formally, the problem can be stated as follows. Given:

- A sequence of input words $w_1...w_m$, for simplicity referred as w.
- The sequence of tags $t_1...t_m$ as t (this way, the set of possible tags is T).
- In our case the context in which a tagging decision is made is represented by the history tuple h: $< t_{-2}, t_{-1}, w_{-2}, w_{-1}, w_0, w_{+1}, sx_0, px_0, cap, num, i >$, where t_{-2} and t_{-1} are the previous two tags, w_{-2} and w_{-1} are

the previous two words at a given position i (this way, H corresponds to the set of all possible histories).

sx and *px* correspond to different sizes of word suffixes, (in this work, x varying from 2 to 4) and prefixes of w_0.

cap and *num* correspond to two binary features to account for capitalization and number status at the current word.

The feature mapping function $\Phi : H \times T \rightarrow \mathbb{R}^d$ maps a history-tag pair to a d-dimensional feature vector we mentioned before. The Structured Perceptron models $P(t|w)$ as $P(t|h;\alpha)$ where $\alpha \in \mathbb{R}^d$ is a parameter vector representing the weight of each feature of Φ. $P(t|h;\alpha)$ is calculated as $\alpha \cdot \Phi(h,t)$ and the objective function is:

$$\hat{t} = argmax_t \quad \sum_1^d \alpha_i \cdot \Phi_i(h,t)$$

Usually the Viterbi algorithm is applied when used on sequence data, in order to efficiently calculate the best tag sequence using dynamic programming. The algorithm is competitive to other options such as maximum-entropy taggers or CRFs [33].

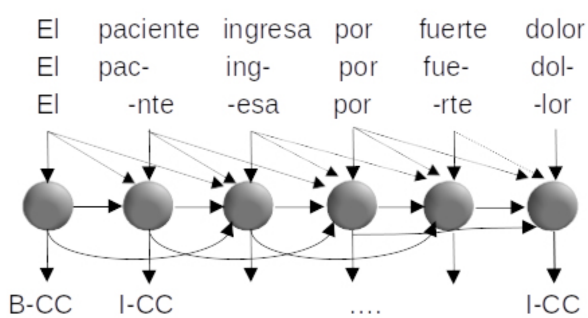


Figure 5: Simplified Architecture of the Structured Perceptron (the upper three rows exemplify the use of word features (first row), 3 letter prefixes and 3 letter suffixes (second and third rows).

We employed our own implementation of this tagger following [37]. We trained 100 iterations and selected the model corresponding to the iteration that achieved the best score on the development set. Although the algorithm achieves a competitive performance compared to state of the art methods, this approach requires a feature engineering effort to identify, select and properly encode relevant features.

3.3.3. Machine Learning: Neural Networks

In addition to a traditional neural network like Perceptron, we explored transfer learning methods. In this case, we used FLAIR [39], a bi-directionally trained Language Model (LM) using Recurrent Neural Networks (RNN), where the basic element is the character and not the word. Based on its characters, FLAIR generates pre-trained contextual embeddings for each word by concatenating the hidden state for the last character of the word in the forward neural network and the first character of the word in the backward neural network, as shown in Figure 6. As described in [39], formally, the objective function of a character-based LM is to maximize the sum of the logs of $P(x_t|x_0, ..., x_{t-1})$, that is to say, an estimate of the predictive distribution over the next character given past characters. FLAIR allows us to combine different types of embeddings by concatenating each embedding vector to form the final word vector. We employed a combination of embeddings as previously reported in section 3.2. One of the main advantages of these methods is that there is no need for feature engineering.

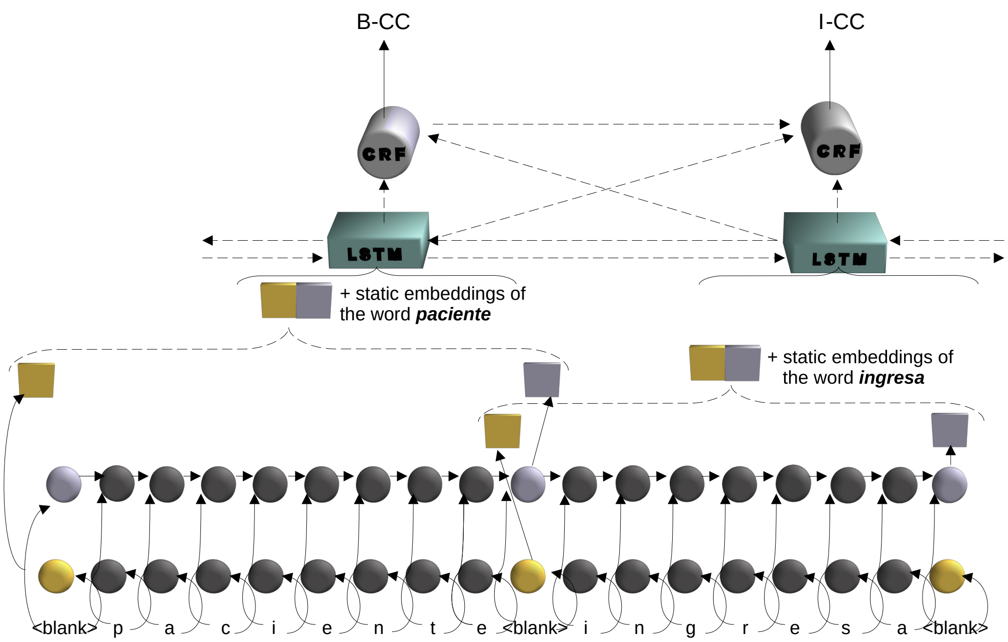


Figure 6: Simplified Architecture of FLAIR (B-CC: start of Chief Complaint, I-CC: continuation of Chief Complaint).

	Micro average			**Macro average (per document)**		
	Precision	Recall	F-score	Precision	Recall	F-score
Rule-based	52.86	51.63	52.24 (14.8)	58.02	57.24	57.62
Perceptron	87.28	85.18	86.22 (3.1)	87.34	86.10	86.72
Neural Networks	93.40	92.55	93.03 (1.8)	91.77	91.26	91.52

Table 5: Results of the different approaches on Section Identification (margin of error with 95% confidence in parentheses).

4. Results

For evaluation, employed the standard measures of precision, recall and F-score defined in formulae (1), where TPS = *correctly identified sections*, FPS = *incorrectly identified sections (marked by the automatic tool and not present in the annotated gold standard)* and FNS = *false negatives, i. e., present in the gold standard and not detected by the automatic tool*. Table 5 shows the main results, given as micro average (over all instances and all documents) and macro average (calculating the mean over the scores on each document).

$$\begin{aligned} Precision &= \frac{TPS}{TPS + FPS} \\ Recall &= \frac{TPS}{TPS + FNS} \\ F{-}score &= \frac{2 * Precision * Recall}{Precision * Recall} \end{aligned} \tag{1}$$

In predicting the section type, the evaluation has been strict in the sense that an error has been counted whenever an automatically detected section did not exactly match with the gold standard section. This was done even when in some cases there is a high degree of overlapping with a correct section (e.g., when the system correctly marks a paragraph as belonging to the Medical History, but it also misses a part of the gold standard section).

Looking at Table 5, we see that the rule-based approach gives the lowest performance (52.24 and 57.62 for micro and macro average, respectively), far from the Machine Learning approaches, and contrary to our first intuition. In general, rule-based solutions tend to have a better precision at the cost of a lower recall, although in this experiment there is not a significant difference between precision and recall. The example rule (1) obtained the best F-score (71) for the *Chief Complaint* section type, and our successive efforts to improve it were not successful, because our attempts to boost recall worsened the precision. Regarding the reason why the rule-based approach gave the worst results, in Information Extraction usually designing more general rules gives an increase in recall, while more specific rules tend to improve precision at the cost of diminishing recall. However, in our particular problem this is not the case, because the objective of the rules is to exactly match entire sections. As a consequence, more general rules have a negative effect on both precision and recall, since incorrectly marking a section produces a cascade effect in the surrounding sections. This was the cause why, after the first successful attempts, dedicating more effort to the rules deteriorated the performance. Thus, we concluded that for this experiment even customized and carefully designed rules with a time-consuming implementation were not able to increase precision. Both ML approaches surpass the performance of the rule-based system, being the neural network based system the best one by a significant margin. The Perceptron-based system outperforms the rule-based one by around 30 absolute points, while neural networks give the best result with 93.03 and 91.52 F-score for micro and macro average, respectively.

Figure 7 presents a detailed comparison of the performance of each approach on the different section types. The rule-based approach presents the lower results for all section types, although the F-score is high in the *Heading* section, which can be considered the easiest to detect. Specially bad is the result for *Exploration*, *Complementary Exploration* and *Evolution*, possibly due to the fact that these sections are not usually marked by an explicit heading. Secondly, the Perceptron based system ranks after the neural network based one, but their F-score is similar for the *Diagnosis* and *Treatment* section types, which can be considered the most important ones from the point of view of the automatic processing of EDSs.

Overall, we can also see how some sections are harder to detect than others. This fact can drastically affect the rule-based system, with big differences according to each section type, and it is related to the difficulty of finding patterns for sections where the headings are absent or also with sections where the headings present a high variability. The differences, albeit smaller, also appear with the Perceptron, which although it is an automatic Machine Learning algorithm, requires an explicit definition of features based on words, suffixes, prefixes or capitalization (feature engineering). In this respect, the diagnoses and treatment sections present the best results, possibly due to the fact that their headings are more predictable. Finally, the neural network system is able to detect the sections without an explicit feature definition. As this system is based on left-to-right and right-to-left vector encodings of the processed text, these systems are able to learn not only from the headings, but also from the vocabulary inside the sections. This is the reason why this system outperforms the other two in the sections with the lowest proportion of explicit headings, like *Exploration* and *Complementary exploration*, where the elements appearing inside the section text, like procedures (ECG, X rays, ...) or clinical measures (sodium, potassium, ...) can help to decide which section is being examined.

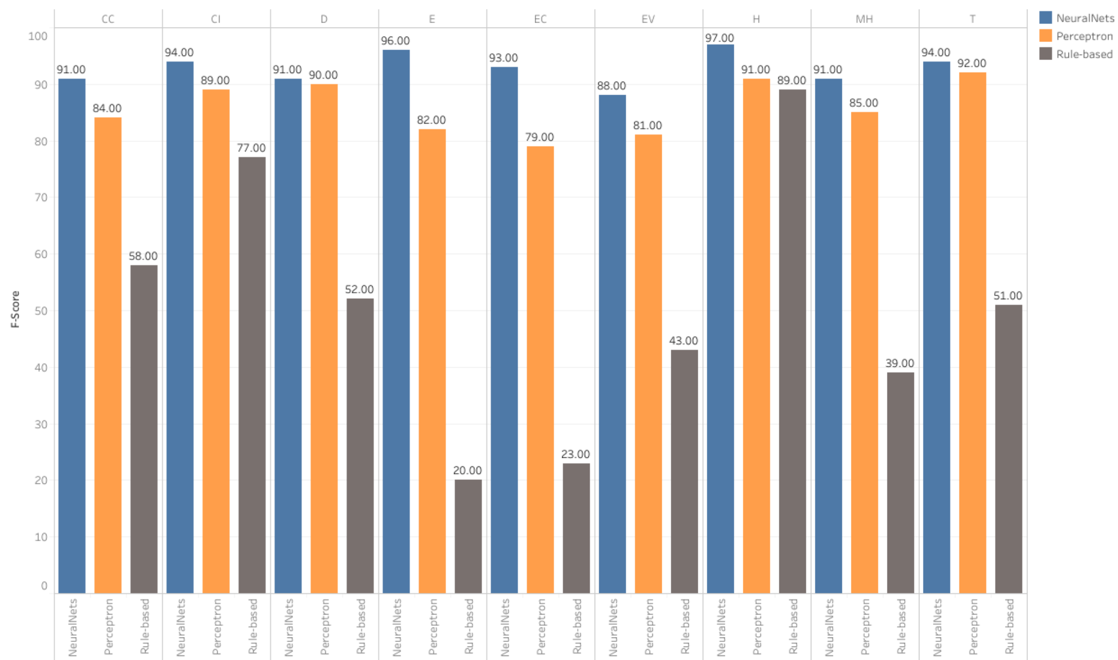


Figure 7: Comparison (F-score) of the three approaches on each section type (H: Heading, CC: Chief Complaint, MH: Medical History, CI: Current Illness, E: Exploration, EC: Complementary Exploration, EV: Evolution, D: Diagnosis, T: Treatment).

Examining the results for the best (neural) system on individual sections, we can see that EV(olution) presents the lowest result (88% F-score), other sections like C(hief) C(omplaint), M(edical) H(istory) and D(iagnosis) give better results (91% F-score), and the remaining sections present higher accuracies.

5. Discussion

In the next subsections we will first (subsection 5.1) look at a new set of experiments to address the effect of applying our system to a new hospital, as usually the writing of EDSs can vary greatly from one hospital to another belonging to the same Health System. Finally, subsection 5.2 presents an analysis of the main error sources.

5.1. Cross hospital generalization

Usually, ML systems tend to obtain good results when the domain of application is the same as the one used for training but, when moving to a different scenario or domain, the results can degrade drastically. We wanted to test the effect of changing the environment of application and, knowing that many times writing styles can vary from one hospital to another, we measured the effect of training using data from one hospital and testing on EDSs from a different hospital. In our case, our data came from two different hospitals from the same hospital system (the Basque Health System, Osakidetza). Figure 8 shows the results when testing on data from a hospital when the training data belongs to the same or a different hospital. The experiments were performed using the best system in Section 4, the neural network based one. As could be expected, the best results are obtained when the training and test sets belong to the same hospital (two left-side bars in each column in Figure 8), and the scores worsen when the system is trained on data from a hospital and applied to the other one (two right-side bars).

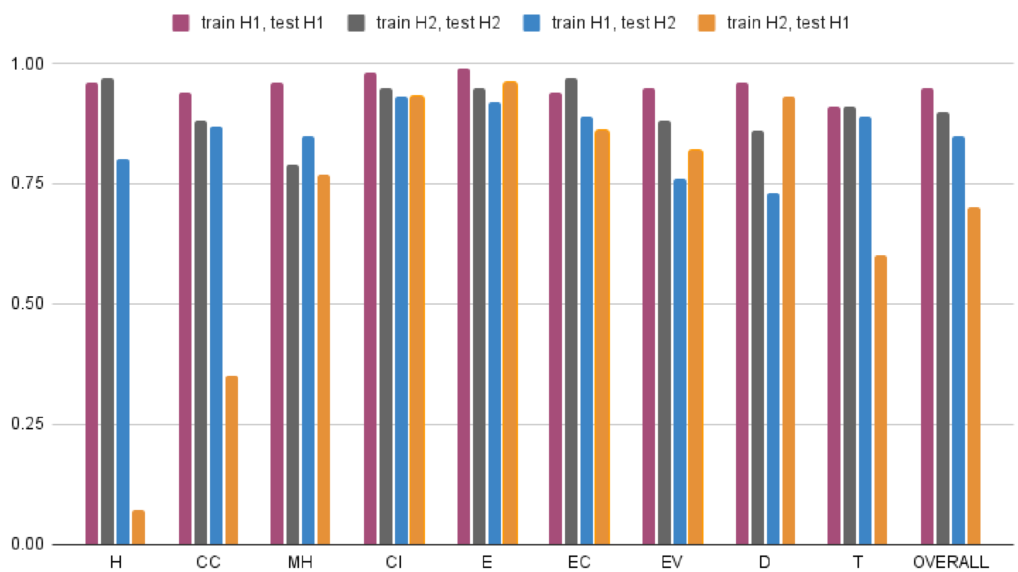


Figure 8: Effect of training and testing on the same or a different hospital (H1: Galdakao-Usansolo hospital, H2: Basurto hospital), measured by F-score.

The difference is significant in almost all types of sections. Specially relevant is the effect of the system trained on hospital H2 and tested on hospital H1 for the Heading section type (H column), where the F-score is very low. We examined the results and concluded that this happens mostly because the headings show a great variation, added to the fact that the data present in the headings is generated automatically most of the times, including record numbers or dates, and this implies that they can be different enough to confuse an automatic system. Surprisingly, this does not happen in the opposite direction, meaning that the data from hospital H1 shows more variability and is useful to account for the instance types of hospital H2. The difference is also significant for the second section type (Chief

Complaint, CC), although less drastic. This was due to a cascade effect as a result of applying a sequence to sequence approach, as the errors in delimiting the first section of the document frequently are carried from one section to the next one. Finally, for some section types, like E(xploration), EV(olution) and D(iagnostic), we can see how applying a system trained on a different hospital can outperform the system based on data from the same hospital. This can be due to the fact that one hospital agrees more with the conventions of the other hospital for these section types.

5.2. Error Analysis

We looked at the errors given by the different systems. For simplicity, we will only examine the results of the best system based on neural networks. An examination of the divergences between the output of the system and the gold standard showed us the main causes of error:

- Errors given by the inherent difficulty of spontaneously written section headings. Although explicit headings are an important clue to delimit sections, the variability of their writings together with the limited size of the training set (100 documents, which means that there are at most 100 instances of each section type) is a source of errors.

- Implicit sections. Some types of sections have a majority of instances without an explicit section heading, which means that the section must be detected using its content words (see Table 2).

- Mixed sections. Although the annotators have decided the exact scope of each section with a high agreement, the use of unstructured and spontaneously written EDSs gives the writers freedom to describe any concept in different places. As an example, the section corresponding to the *Medical History* can contain passages related to past diagnoses, treatments and explorations, which can pose a challenge for an automatic system.

- A special case of mixed sections can be the confusion between two related section types:
  - *Chief Complaint* and *Current Illness*. These two sections present the most diffuse definition [27], and are the cause of several errors.
  - *Exploration* and *Complementary Exploration*. Although the definition of each section is precise, sometimes physicians mix them in the same block or paragraph.

In Section 3.1.2, we mentioned that the ordering of section types shows a great variability. In order to measure its impact on the results, we split the test set in two subsets. The first subset corresponds to the documents that follow the canonical order (26% of the documents), while the rest of the documents conform the second subset (non-canonical order and/or missing sections, 74% of the documents). Since our sequence learning-based methods depend on the ordering for predicting the next token, this has an effect in the IOB-labeling prediction, with a F-score of 95.40 for the canonical documents and 89.81 for the non-canonical ones.

Figure 9 presents the main types of mistakes made by the automatic tool. It shows how the errors are concentrated in some sections, like Chief Complaint (CC), Medical history (MH) and Diagnosis (D). Overall, the distinction of different sections is reflected in the text by means of different clues, ranging from semantics (the content of each section) to syntax (e.g., use of section headings and separate paragraphs or text blocks for each section) but, in most of the errors, these conventions do not hold, and this causes the automatic tool to find an additional difficulty.

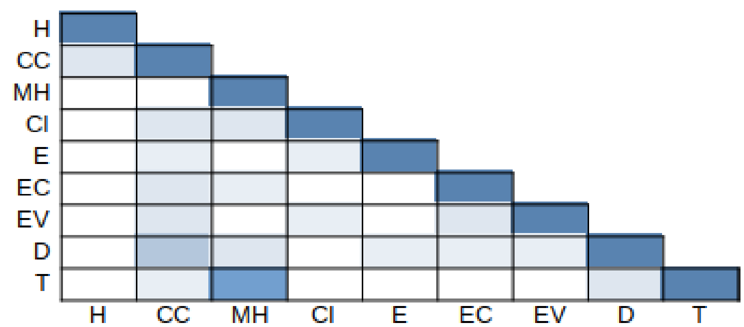


Figure 9: Confusion matrix, where darker green means a higher frequency, of each instance (H: Heading, CC: Chief Complaint, MH: Medical History, CI: Current Illness, E: Exploration, EC: Complementary Exploration, EV: Evolution, D: Diagnosis, T: Treatment).

6. Conclusion

We present a system for Section Identification in Discharge Summaries written in Spanish. We have adopted an annotation model based on H7 CDA R2 for Electronic Discharge Summaries (EDS) of the Spanish Health System, and we have applied it to manually annotate a corpus of 300 EDSs, obtaining a high inter-annotator agreement.

We have evaluated the contribution of different rule-based and Machine Learning approaches and study the strengths and weaknesses of each option. Most previous works have used section identification as an auxiliary module for carrying on clinical processing, relying on a rule-based approach. However, our results show that section identification is a task on its own, where simple methods do not obtain the best results. The Machine Learning systems obtain results that are good enough for the application of the system in a production setting. Specifically, we show that Language Model tuning is a key factor, as a Language Model-based transfer learning provides the best performance. The paper has also studied the generalization ability of models trained in different hospitals, showing that different section types have significant differences in some cases.

The developed automatic annotation models and software are freely available contacting the authors.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31), the European Commission (FEDER), the Basque Government (IXA IT-1343-19), and the EU ERA-Net CHIST-ERA and the Spanish Research Agency (ANTIDOTE PCI2020-120717-2).

References

[1] C. Peterson, C. Hamilton, P. Hasvold, From innovation to implementation – eHealth in the WHO European Region, World Health Organization, 2016.

[2] M. Adnan, J. Warren, M. Orr, A. Ewens, J. Scott, S. Trubshaw, The quality of electronic discharge summaries for post-discharge care: Hospital panel assessment and it to support improvement, Health Care and Informatics Review Online 15 (2011).

[3] Health at a Glance: Europe 2018 STATE OF HEALTH IN THE EU CYCLE, https://ec.europa.eu/health/sites/default/files/state/docs/2018_healthatglance_rep_en.pdf, 2020. Last Online; accessed 31-05-2021.

[4] Health IT Data Summaries, https://dashboard.healthit.gov/apps/health-information-technology-data-summaries.php, 2021. Last Online; accessed 31-05-2021.

[5] Connecting health and care for the nation: A shared nationwide interoperability roadmap. Office of the National Coordinator for Health Information Technology (ONC). Washington, DC: U.S. Department of Health and Human Services (HHS), 2015.

[6] Recommendation on a European Electronic Health Record exchange format. European Commision, 2019.

[7] State of Interoperability among U.S. Non-federal Acute Care Hospitals in 2018 , https://www.healthit.gov/sites/default/files/page/2020-03/State-of-Interoperability-among-US-Non-federal-Acute-Care-Hospitals-in-2018.pdf, 2020. Last Online; accessed 31-05-2021.

[8] openehr, https://www.openehr.org, 2020. Last Online; accessed 31-05-2021.

[9] Health Level Seven (HL7). FHIR, `http://www.hl7.org`, 2019. Last Online; accessed 31-05-2021.

[10] Health Level Seven (HL7). CDA, `http://www.hl7.org`, 2019. Last Online; accessed 31-05-2021.

[11] H. K., K. Saranto, N. P., Definition, structure, content, use and impacts of electronic health records: a review of the research literature, Int J Med Inform. 77 (5) (2008) 291–304.

[12] W. LL., Medical records that guide and teach, N Engl J Med. 14) (1968) 593–600.

[13] A. Pomares-Quimbaya, M. Kreuzthaler, S. Schulz, Current approaches to identify sections within clinical narratives from electronic health records: a systematic review, BMC Medical Research Methodology 19 (2019).

[14] T. Edinger, D. Demner-Fushman, A. Cohen, S. Bedrick, H. W., Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval, in: AMIA Annu Symp Proc., pp. 660–669.

[15] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, A comprehensive study of named entity recognition in chinese clinical text, Journal of the American Medical Informatics Association 21 (2014).

[16] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: A literature review, Journal of Biomedical Informatics 77 (2018) 34 – 49.

[17] H.-J. Lee, Y. Zhang, M. Jiang, J. Xu, C. Tao, H. Xu, Identifying direct temporal relations between time and events from clinical notes, BMC Medical Informatics and Decision Making 18 (2018).

[18] A. Pérez, K. Gojenola, A. Casillas, M. Oronoz, A. Díaz de Ilarraza, Computer aided classification of diagnostic terms in spanish, Expert Systems with Applications, **42**(6), 2949–295 (2015).

[19] A. Atutxa, A. D. de Ilarraza, K. Gojenola, M. Oronoz, O. P. de Viñaspre, Interpretable deep learning to map diagnostic texts to icd-10 codes, International Journal of Medical Informatics 129 (2019) 49 – 59.

[20] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A. K. Khanna, J. B. Cywinski, K. Maheshwari, P. Xie, E. P. Xing, Multimodal machine learning for automated icd coding, in: F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), Proceedings of the 4th Machine Learning for Healthcare Conference, volume 106 of *Proceedings of Machine Learning Research*, PMLR, Ann Arbor, Michigan, 2019, pp. 197–215.

[21] A. Duque, H. Fabregat, L. Araujo, J. MartinezRomo, A keyphrasebased approach for interpretable ICD-10 code classification of Spanish medical reports, Artificial Intelligence in Medicine (2020).

[22] S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers, A. Löser, Sector: A neural model for coherent topic segmentation and classification, Transactions of the Association for Computational Linguistics 7 (2019) 169–184.

[23] E. Choi, Z. Xu, Y. Li, M. W. Dusenberry, G. Flores, E. Xue, A. M. Dai, Learning the graphical structure of electronic health records with graph convolutional transformer, in: Association for the Advancement of Artificial Intelligence (AAAI).

[24] S. Rosenthal, K. Barker, Z. Liang, Leveraging medical literature for section prediction in electronic health records, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4864–4873.

[25] E. Rush, I. Danciu, G. Ostrouchov, K. Cho, B. Mayer, Y.-L. Ho, J. Honerlaw, L. Costa, F. Linares, E. Begoli, Jsonize: A scalable machine learning pipeline to model medical notes as semi-structured documents, AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2020 (2020) 533–541.

[26] L. K. Branting, C. Pfeifer, B. Brown, L. Ferro, J. Aberdeen, B. Weiss, M. Pfaff, B. Liao, Scalable and explainable legal prediction, Artificial Intelligence and Law (2020).

[27] A. R. Terroba, Mejora de la calidad del informe clínico de alta hospitalaria desde el punto de vista lingüístico, PhD Thesis, University of La Rioja (2018).

[28] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, CoRR abs/1301.3781 (2013).

[29] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

[30] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

[31] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Takefuji, Wikipedia2vec: An optimized tool for learning embeddings of words and entities from wikipedia, arXiv preprint 1812.06280 (2018).

[32] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649.

[33] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 188–191.

[34] A. Jagannatha, H. Yu, Bidirectional recurrent neural networks for medical event detection in electronic health records, CoRR abs/1606.07953 (2016).

[35] A. Atutxa, K. Bengoetxea, A. D. de Ilarraza, M. Iruskieta, Towards a top-down approach for an automatic discourse analysis for basque: Seg-

mentation and central unit detection tool, PLoS ONE 14(9): e0221639 (2019).

[36] H.-J. Dai, S. Syed-Abdul, C.-W. Chen, C.-C. Wu, Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields, BioMed Research International 2015 (2015).

[37] M. Collins, Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Association for Computational Linguistics, 2002, pp. 1–8.

[38] A. Pérez, R. Weegar, A. Casillas, K. Gojenola, M. Oronoz, H. Dalianis, Semi-supervised medical entity recognition: A study on spanish and swedish clinical corpora, Journal of Biomedical Informatics 71 (2017).

[39] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 54–59.