applied sciences

MDPI

Article

Integrating Speculation Detection and Deep Learning to Extract Lung Cancer Diagnosis from Clinical Notes

Oswaldo Solarte Pabón [1,2,*], Maria Torrente [3], Mariano Provencio [3], Alejandro Rodríguez-Gonzalez [1,2] and Ernestina Menasalvas [1,2]

[1] Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón, Madrid, Spain; alejandro.rg@upm.es (A.R.-G.); ernestina.menasalvas@upm.es (E.M.)
[2] Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain
[3] Hospital Universitario Puerta de Hierro, 28222 Majadahonda, Madrid, Spain; maria.torrente@salud.madrid.org (M.T.); mariano.provencio@salud.madrid.org (M.P.)
* Correspondence: oswaldo.solartep@alumnos.upm.es

Abstract: Despite efforts to develop models for extracting medical concepts from clinical notes, there are still some challenges in particular to be able to relate concepts to dates. The high number of clinical notes written for each single patient, the use of negation, speculation, and different date formats cause ambiguity that has to be solved to reconstruct the patient's natural history. In this paper, we concentrate on extracting from clinical narratives the cancer diagnosis and relating it to the diagnosis date. To address this challenge, a hybrid approach that combines deep learning-based and rule-based methods is proposed. The approach integrates three steps: (i) lung cancer named entity recognition, (ii) negation and speculation detection, and (iii) relating the cancer diagnosis to a valid date. In particular, we apply the proposed approach to extract the lung cancer diagnosis and its diagnosis date from clinical narratives written in Spanish. Results obtained show an F-score of 90% in the named entity recognition task, and a 89% F-score in the task of relating the cancer diagnosis to the diagnosis date. Our findings suggest that speculation detection is together with negation detection a key component to properly extract cancer diagnosis from clinical notes.

Keywords: Natural Language Processing (NLP); information extraction; diagnosis extraction; lung cancer; deep learning; speculation detection; negation detection

Citation: Solarte Pabón, O.; Torrente, M.; Provencio, M.; Rodríguez-Gonzalez, A.; Menasalvas, E. Integrating Speculation Detection and Deep Learning to Extract Lung Cancer Diagnosis from Clinical Notes. *Appl. Sci.* **2021**, *11*, 865. https://doi.org/10.3390/app11020865

Received: 18 December 2020
Accepted: 12 January 2021
Published: 19 January 2021

1. Introduction

The steady growth in the adoption of Electronic Health Records (EHR) around the world has introduced the possibility of extracting hidden information from clinical notes [1]. These notes contain useful and valuable information to support clinical decision making [2,3]. However, the information in clinical notes is presented in a narrative form, which makes the task of structuring the data especially challenging. Extracting this information manually would not be a viable task because it would be time-consuming and costly. Although Natural Language Processing (NLP) tools aim to automatically extract medical concepts, once these concepts are extracted, one must relate them to dates and ensure they are not speculated or negated.

In recent years, deep learning-based approaches have shown to be effective in improving the extraction of medical concepts from clinical narratives [4,5]. However, there is still a gap between concept extraction and concept understanding [6]. Extracting medical concepts is not enough when it comes to understanding events concerning a patient, additional steps are therefore required [6].

Lung cancer is one of the most common chronic diseases in the world and the leading cause of cancer death among both men and women (https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.

html). Therefore, the accurate identification of lung cancer information is crucial to support clinical and epidemiological studies [4,7]. The cancer diagnosis is an important step for both the effective control of the disease, and the development of treatment plans [8]. Data generated during both clinical care processes and research in oncology have increased considerably in recent years [9].

In [10] a pipeline to extract lung cancer diagnosis from clinical notes written in Spanish was proposed. Although this proposal showed promising results, it has two main weaknesses: (i) a limitation to relate a cancer concept with a proper diagnosis date, and (ii) a limitation to properly recognize cancer concepts that have been affected by speculation.

Speculation is a linguistic phenomenon that frequently appears in clinical notes, and is inherent in many medical decisions [11]. Moreover, in the medical domain, some authors also refer to speculation as uncertainty [12,13]. Physicians often face uncertain findings when diagnosing or treating patients. Therefore, for text mining applications, it is important to detect uncertainty because uncertain findings can be incorrectly identified as real or factual events [14].

In this paper, we propose an approach that extends and improves the proposal presented in [10]. This approach combines deep learning-based and rule-based methods to address the previously mentioned limitations. The proposed approach aims to automatically extract the lung cancer diagnosis from clinical notes and consists of three steps: (i) lung cancer named entity recognition using deep-learning approaches, (ii) speculation and negation detection and, (iii) relating the cancer diagnosis to a proper date.

The remaining sections of this research paper have been organized as follows: Section 2 reviews the most recent studies on cancer concept extraction from clinical narratives. Section 3 shows the proposed hybrid approach to extract lung cancer diagnosis. Section 4 describes a deep-learning approach used to extract lung cancer concepts. Section 5 deepens on the rule-based approach for speculation and negation detection, while in Section 6 the process to relate the cancer concept to a diagnosis date is explained. Section 7 presents the validation and results and Sections 8 and 9 present the discussion and final concluding remarks, as well as an outlook for future studies.

2. Related Work

Clinical concept extraction refers to automatically extracting concepts of interest from unstructured clinical texts [15]. Concept extraction can also be referred to as Named Entity Recognition (NER) [16]. In the cancer domain, research to extract cancer concepts from clinical narratives has increased considerably in recent years due to its benefits associated to evidence-based research and quality improvement [17]. Clinical concept extraction is commonly addressed by using rules, machine learning, or deep learning approaches.

One of the first interests in the field of oncology was identifying a patient's cancer stage, which is an important prognostic factor in order to understand cancer-survival. In [18,19], a rule-based method to extract data associated to a tumor's stage from clinical documents of lung cancer patients is presented. Meanwhile, in [20], a different rule-based tool for extracting cancer staging from pathological reports of breast cancer using the TNM (https://www.cancer.gov/about-cancer/diagnosis-staging/staging) notation is described. Other studies related to the extraction of a cancer stage are described in [21–23]. Although the stage is an important cancer factor, extracting only this concept is not enough to fully understand cancer behavior.

Other proposals have attempted to extract more concepts associated with cancer. These approaches focus on the extraction of cancer diagnosis [9,24], treatments [25,26], or both [4,5]. In [27], the authors proposed a strategy to extract breast cancer diagnosis, the histology of malignant neoplasm, temporal expressions, and recurrent cancer-related events. In [25], a machine learning approach to extract cancer treatments, doses, toxicities, and dates from lung cancer patients is proposed. Other machine learning-based approaches to extract cancer information are proposed in [28,29].

Recently, Deep neural networks have been shown to improve performance in the processing of natural language texts. More specifically, Recurrent neural networks (RNN) [30–32] and Convolution Neural Networks (CNN) [33] have been used to improve text processing. Deep learning approaches have also shown their applicability in different domains such as finance [34], smart cities [35], and security [36]. Besides, word embedding is a useful tool to create deep learning applications [37–39], and it has gained great popularity in the biomedical field [40,41].

In [4] a deep learning-based approach to extract lung cancer stages, histology, tumor grades and therapies (chemotherapy, radiotherapy, surgery) is proposed. Clinical notes, pathology reports, and surgery reports are used to test the system. The authors highlight the feasibility of extracting cancer-related information from clinical narratives and the feasibility of improving the efficiency of humans through NLP techniques. A Bidirectional Long Short-Term Memory (BiLSTM) network is used in [5] to extract a comprehensive set of breast cancer concepts. This proposal extracts more than forty concepts which help to understand the behavior of this disease.

Although deep learning-based approaches have improved the ability to extract medical concepts, there is still a gap between concept extraction and concept understanding [6]. Extracting cancer-related concepts alone is not effective in understanding clinical events related to patients. New steps are required to go beyond clinical concept extraction in order to understand relationships between events and concepts in the natural history of a patient.

In addition, most of the proposals mentioned above have focused on the English language. According to [42], information extraction in the medical domain also represents its own challenges in languages other than English. In the Spanish language, a rule-based approach to extract concepts such as stage, performance status, and mutations in the lung cancer domain is proposed in [8]. To deal with the recognition of time expressions, in [43] a temporary Tagger Annotator is proposed. This tool aims to identify and normalize time expressions in clinical texts written in Spanish. However, extracting cancer concepts and temporal expressions alone is not enough and, these concepts have to be related [44].

Deep learning approaches have also been proposed to extract cancer-related concepts in clinical texts written in Spanish [45–47]. However, one limitation of these proposals is that they only extract cancer entities. Cantemist [48], an annotated corpus containing only one labeled entity (Cancer tumor morphology) is used to extract the cancer concepts. To the best of our knowledge, there is no approach to extract cancer concepts and relate them to other concepts such as dates for the case of clinical notes written in the Spanish language.

3. A Hybrid Approach to Extract Lung Cancer Diagnosis

When a patient goes to a hospital, different tests are performed to confirm the final diagnosis. This process involves numerous interactions between doctor and patient. Consequently, different notes are generated in which the doctor reports the patient antecedents, physical status, and the suspected diagnosis. When the patient is finally diagnosed, more narratives will be generated with the diagnosis, date, treatments and all information of the the follow up.

Figure 1 shows fragments of clinical narratives for a patient suffering lung cancer. In the text one can observe that there are different ways to refer to cancer diagnosis concepts and distinct ways to mention dates. The correct identification of the diagnosis date is a challenge since many date annotations can be extracted. Moreover, the following has to be taken into account for the proper identification of the diagnosis:

- Some cancer concepts are associated with a family member, but not the patient.
- Speculation and negation are two linguistic phenomena that frequently appear in clinical texts.
- Different events can occur to the patient (begin clinical trial, diagnosis, start treatment, etc.). All these events will appear with a date and can appear in a sentence in which a mention to a cancer concept is present. So it is needed to distinguish the events and the dates. As an example, let's take the sentence: *"Patient diagnosed with lung cancer,*

treated with surgery in March 2017"; in this case the date refers to the surgery and not to the diagnosis. Thus apart from cancer and date annotations, an approach to find events and to related dates to events is required.

- A medical record can contain hundreds of clinical notes for the same patient; this fact represents a challenge to extract the correct diagnosis date automatically. This problem is enhanced when a hospital is treating thousands of cancer patients, and therefore the number of clinical notes grows rapidly.

Thus, to deal with this challenging task, we propose a hybrid approach that combines deep learning-based and rule-based methods. The main goal of this proposal is to improve the lung cancer diagnosis extraction process. Figure 2 shows the proposed approach that integrates three steps: (i) Lung cancer named entity recognition, (ii) Negation and speculation detection, and (iii) Relating cancer diagnosis and dates. In the next sections, we will describe each step.

Family History: Father with lung cancer in September 2001.

Biopsy test: ***Negative*** for cancer, 03/10/2016.

Lung carcinoma is ***probable***, but I will wait for TAC results, February 2017.

12/03/2017 The test ***does not*** confirm lung cancer.

Patient with ***possible*** lung neoplasm, 2017/03/21 .

(25-05-2017) TAC , Patient with Small cell lung carcinoma .

Carcinoma treated in October 2018 .

Patient **diagnosed** with a carcinoma, ***begins treatment*** with chemotherapy in November 2017.

Lung carcinoma stage T2N0M0, **treated** with lobectomy in July 2017.

Figure 1. Medical history fragments.

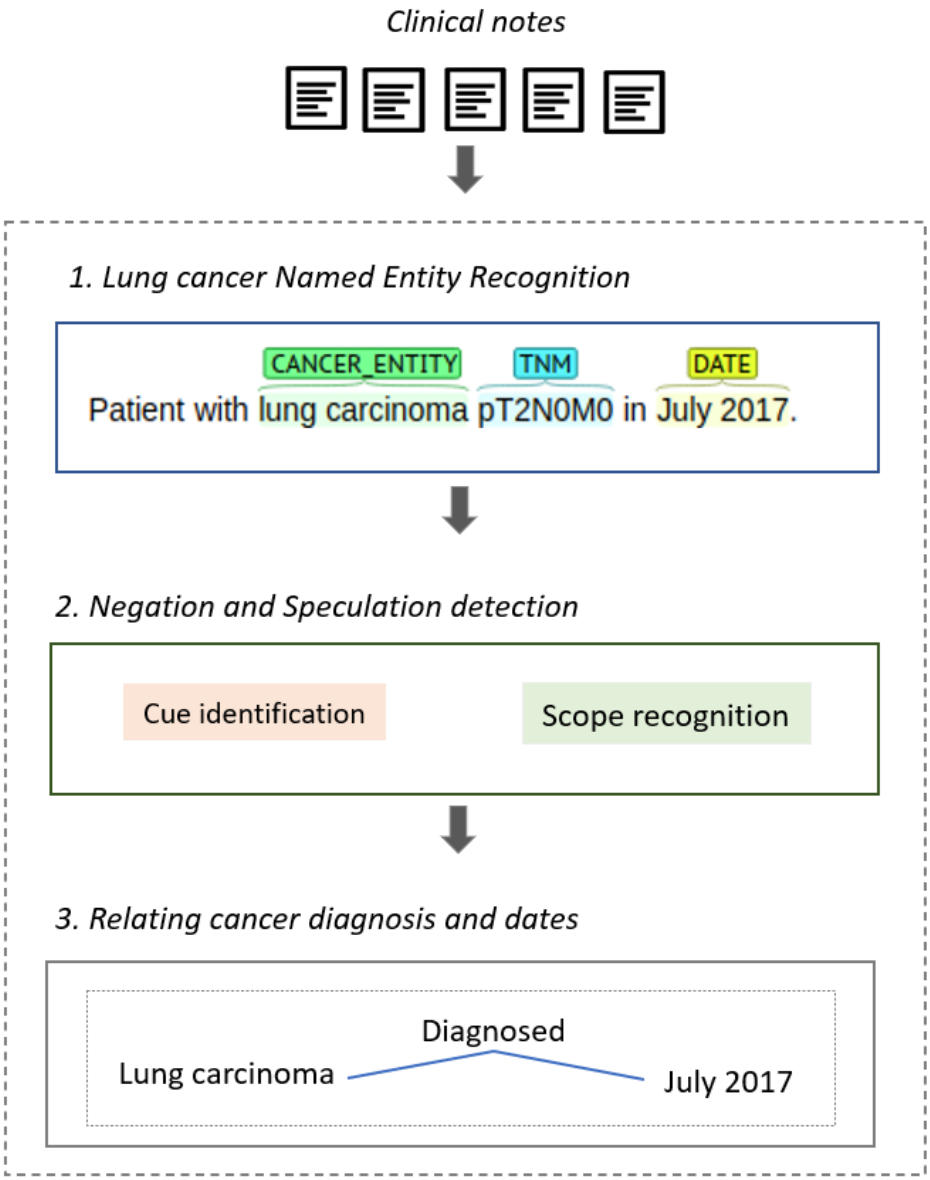

Figure 2. Approach to extract lung cancer diagnosis.

4. Lung Cancer Named Entity Recognition

In this section, a deep learning model to extract lung cancer named entities from clinical notes written in Spanish is described. The main goal of this model is to improve lung cancer concept extraction. To create this model, a corpus with the entities to be extracted was first annotated, followed by a trained BiLSTM (Bidirectional Long Short-Term Memory) neural net to carry out named entity recognition.

4.1. Corpus Annotation

The corpus has been manually annotated using the Brat (https://brat.nlplab.org/) tool. It is composed of clinical notes from lung cancer patients from *"Hospital Universitario Puerta de Hierro, Madrid, Spain"*, with every clinical note being introduced anonymously. The corpus contains the following labeled entities:

- **Cancer entity:** this entity captures both the cancer type *(carcinoma, adenocarcinoma, cancer, etc.)* together with the anatomical location *(left lung, right lung, right lobe, etc.)*. For instance: *"Patient diagnosed with right lung adenocarcinoma"*. In the annotation process the more detailed description will be used. Thus in the sentence, *"Patient*

diagnosed with small cell lung cancer", the concept "Small cell lung cancer" will be annotated instead of *"lung cancer"*.

- **Cancer stage:** staging is the process of finding out how much cancer is in a person's body and where it's located. The cancer stage can be expressed on a scale that ranges from I to IV. Stage I indicates the initial stage and stage IV, the most advanced stage of cancer. The cancer staging can also be done using the TNM notation: Tumor (T), Nearby(N), and Metastasis(M). As a consequence, we have annotated both scales. On the one hand, entities such as *stage II and stage IV* are annotated as a stage entity while expressions such as *cT3cN3cM1* and *T3 N2 M0* are annotated as TNM entity.
- **Dates:** Represents dates and time expressions mentioned in clinical notes. Date entity is a crucial concept to obtain the natural history of the patient. Only explicit dates are annotated.
- **Events**: This entity is used to represent events such as *being diagnosed, being treated, treatment start, begin clinical trial, etc.*. In the sentence *"Patient **diagnosed** with lung cancer, **begins treatment** with chemotherapy on 5 December 2019*, there are two events (shown in bold).
- **Family members**: represents concepts about family members of a patient. This entity commonly appears together with cancer concepts in the family antecedents, *(e.g., "**Mother** diagnosed with lung cancer in 2007)"*. This entity is used to differentiate between a cancer concept associated to a patient, and one associated to a family member.
- **Treatment**: The kind of treatment ranging from chemotherapy, radiotherapy, and surgery will be included. These treatments will be annotated separately. This tag is only used to annotate generic mentions to treatments, without mentioning the specific drug, as in the case of sentences such as: *"Patient with lung cancer, with **chemotherapy** in October 2018"*.
- **Drug**: This entity is used to annotate particular names of drugs related to any kind of treatment, such as in the sentence: *Lung cancer patient, treated with **Carboplatin** on October 2018*.

Table 1 shows a summary of the number of annotations for each entity. Two domain experts annotated the corpus previously described. The inter-annotator agreement was 0.89 measured by Cohen's Kappa.

Table 1. Number of annotations for each entity in the corpus.

Entity	Number of Annotations
Cancer entity	4128
Stage	3274
TNM	1152
Date	4312
Family member	883
Events	2703
Treatment	1302
Drug	463

4.2. A Deep Learning Approach to Extract Lung Cancer Entities

The proposed model (see Figure 3) uses a BiLSTM neural net with a sequential CRF layer. The model is based on the architecture proposed by Lample et al. [32]. The model takes sentences with a vector representation as the input, and the output is the predicted label for each word in a sentence. The BIO tagging format is used to represent predicted entities, labelling each entity with: B (at the beginning of the entity), I (inside the entity), or O (Outside the entity). Figure 3 depicts this model, which consists of three layers: (i) Embedding layer, (ii) BiLSTM layer, and (iii) CRF layer.

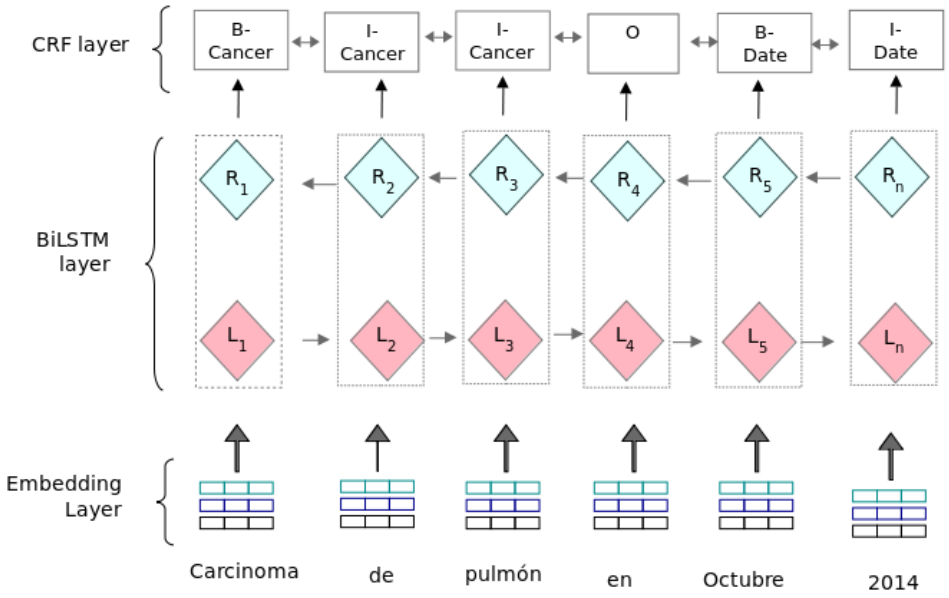

Figure 3. Bidirectional Long Short-Term Memory/Conditional Random Field (BiLSTM-CRF) model.

- **Embedding layer**: this layer makes it possible to represent words and documents using a dense vector representation. Word embeddings allows words with similar meanings to have a similar representation. The model that is proposed has been trained with the inclusion of different biomedical embedding:
  - SciELO Full-Text: This word embedding was created using full-text medical articles from Scielo, a scientific electronic library [41].
  - WikiHealth: This embedding was generated using a subset of Wikipedia articles comprised by the categories of Pharmacology, Medicine and Biology [41].
  - Lemma and Part of Speech (POS): We create in-house embeddings with lemmas and POS tags using as input the sentences in the annotated corpus.

- **BiLSTM layer:** as the input setting, this layer uses different embeddings and processes each vector representation of the text sequence in two ways:
  - A forward computation process each sequence from left to right. In Figure 3, each L_i represents the value of left context for the word n in the sequence.
  - A backward computation process from right to left. In Figure 3, each R_i represents the value of the right context for the word n in the sequence.

  As an output, this layer produces a vector representation for each word, concatenating the left and right context values. These vectors contain scores for each label that is to be predicted. In the BiLSTM model, each input sentence is contextualized both on the left (L) and on the right (R). Both LSTM (left and right) are independent but contextualized with the same distribution function.

- **CRF layer:** This layer decodes the best label in all possible labels using the CRF (Conditional Random Fields) algorithm proposed by [49]. This algorithm considers the correlations between other labels and jointly decodes the best chain of labels for a given input sentence of text. Although the BiLSTM layer produces scores for each label, these scores are conditionally independent. For sequence labeling tasks, it is necessary to consider the correlations between labels. Therefore, the CRF layer aims to model dependencies between these labels to improve the predictions for each label. As an input setting, the CRF layer takes the output vectors from the BiLSTM layer and outputs the label sequence with the highest prediction score.

5. Negation and Speculation Detection

The deep learning approach described previously has been able to extract cancer-related entities. However, many of these entities can be affected by negation or speculation. Detecting speculation and negation is a crucial step to extract a cancer diagnosis correctly. This section shows a rule-based approach to detect negation and speculation in clinical texts written in Spanish. This task is commonly divided into two subtasks: *Cue identification* and *Scope recognition*. Cues are words or terms that express negation (e.g., not, nothing, negative) or speculation (e.g., possible, probable, suggest) [13]. The scope is the text fragment affected by a cue in a sentence [50]. In the next example, the cue is shown in bold, and the scope is underlined.

"Patient with ***possible*** *lung cancer in July 2014, we recommend a chest test to confirm."*.

Given a text sentence, the cues will first be detected, and later the tokens affected by these cues will be found. We propose to enrich this process by improving the lexicon used with popular medical cues. Besides, an analysis of the sentences in which negation or speculated terms appear is performed to improve scope recognition.

The main motivation to enrich the negation and speculation detection is that clinical records are written by highly skilled physicians and nurses using domain-specific terms. These records are written under time pressure, the text is short and efficient, and written in telegraphic style [51]. Consequently, clinical texts can contain sentences shorter than those found in other domains, and the lexicon to express speculation and negation can be richer and more complex.

Taking into account the above facts, this approach is composed of the following steps: (i) developing a cue lexicon focused on the medical domain, (ii) defining regular expressions for detecting cues in clinical text, (iii) analyzing the sentences where the cues are detected and finally, (iV) recognizing the scope for each detected cue.

5.1. Developing a Cues Lexicon

This step aims to develop a lexicon that contains negation and speculation cues in Spanish, enriched with popular cues in the medical domain. The strategy shown in Figure 4 was applied for developing the lexicon. This strategy consists of two steps: Cue Collection and Semantic Enrichment.

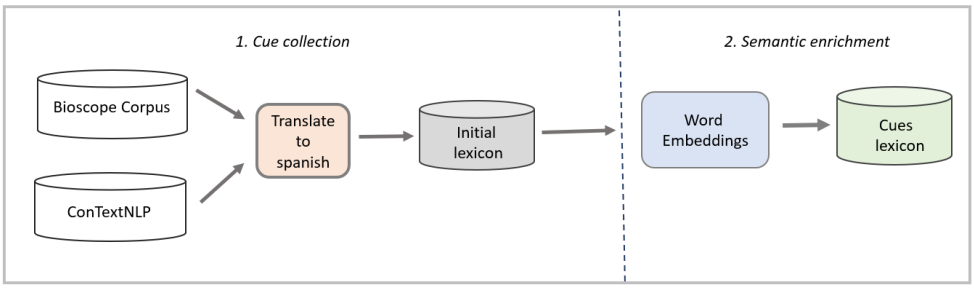

Figure 4. Strategy to generate the cues lexicon.

- **Cue Collection:** The main goal of this step is to get an initial lexicon containing an initial set of cues in Spanish. Two public resources are used: BioScope corpus [52], and ConText [53]. The collected cues are translated to Spanish using Google translate, and later they are manually corrected. Bioscope and ConText are chosen as these resources are frequently used for negation and uncertainty detection in the English language.
- **Semantic Enrichment:** this step aims to enrich the initial lexicon by adding new cues that are semantically related. For each cue in the initial lexicon, similar words indicating negation or speculation are found. The semantic enrichment step is performed using the word embedding technique [38,54]. Since the word embedding technique generates high-dimensional vectors for each word, the t-distributed stochastic neigh-

bor embedding (t-SNE) technique [55] was used to visualize the similarity between words using only two dimensions. Figure 5 shows in a two-dimensional diagram a set of words (*printed in blue*) semantically related to the cue *"Probable"* (*printed in red*), and other terms (*printed in green*) that appear within the context of this cue. The initial lexicon contains 390 cues, and after the semantic enrichment step, the cue's lexicon contains 512 cues.

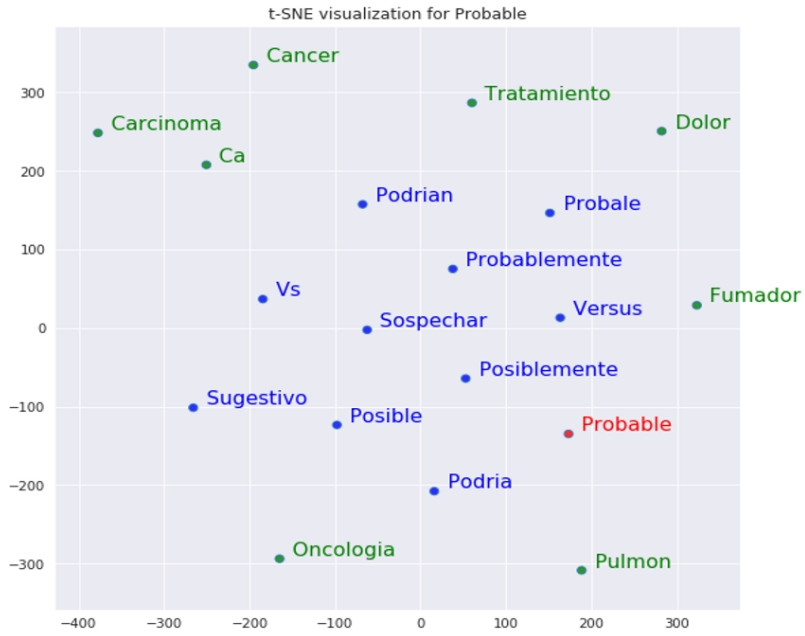

Figure 5. Related words to the cue "Probable".

5.2. Cue Detection in Clinical Texts

This step detects speculation and negation cues in a clinical text. Cue detection receives two inputs: a text sentence and the cues lexicon (Figure 4). To perform this task four regular expressions (*Regex)* are used. The first two Regex were adapted from the Negex proposal [56]. The next two Regex were adapted from [57] to improve multiple and contiguous cue detection. In our approach, those regular expressions are adapted for detecting both negation and speculation.

1. <Prefix-cue> * <UMLS terms>
2. <UMLS terms> * <Postfix-cue>

The symbol * represents an unspecified number of tokens in the sentence. UMLS (*Unified Medical Language System*, https://www.nlm.nih.gov/research/umls/index.html) terms represent medical concepts or findings affected by a cue. The cues are divided into two groups depending on their location with respect to the terms they affect:

- **Prefix-cue:** is found before the affected tokens. For example: in the sentence:

  "Paciente con ***probable*** *carcinoma pulmonar". (Patient with probable lung carcinoma)*, the cue is the word *"probable"* and the tokens affected are underlined. In these cases the tokens affected appear to the right of the cue.

- **Postfix-cue:** is found after the affected tokens. In the next sentence the cue is *"negativa"* and the tokens affected are underlined. In these cases the tokens affected appear to the left of the cue:

  "Biopsia líquida para cáncer: ***negativa****". (Liquid biopsy for cancer*: negative).

The third regular expression is used for detecting multiple contiguous Prefix-cues. When a cue appears multiple times contiguously is considered a contiguous-cue [58]. For example: the sentence,

*"**No** dolor, **no** vómitos, **ni** fiebre, **no** tos"* . (No pain, no vomiting, no fever, no cough), contains four contiguous cues. The affected tokens are underlined.

3. <Prefix-contiguous-cue> * <UMLS terms>
4. <UMLS terms> * <Postfix-contiguous-cue>

The fourth regular expression is used to detect multiple contiguous Postfix-cues. In the next example, there are two contiguous cues and the affected tokens are underlined.

"Análisis de orina: ***Negativo****, glóbulos rojos:* ***negativo****."* (Urinalysis: Negative, Red Blood Cells: Negative).

5.3. Sentence Analysis

Previous rule-based approaches [53,56,58] use a stop word to find the negation or speculation scope. However, it is important to analyze the sentence in which the cue appears in order to improve rules. In particular, analyzing the sentence's length and the presence of contiguous cues in order to define rules for scope recognition is proposed:

- *Sentence length:* is the number of tokens in the sentence in which speculation or negation cues were detected. This property aims to analyze the behavior of negation and speculation according to the sentence's length. The sentence length is used to define the *short sentence heuristic*. This heuristic indicates that a sentence is considered as a *short sentence* if its length belongs to the first quartile, and it has only one cue. Although short sentences are more frequently found in clinical text [51], long sentences can also be found [57]. Therefore, the short sentence heuristic is used to choose the proper rule in the scope recognition step according to sentence length.
- *Presence of contiguous cues:* this aims to analyze the behavior of negation and speculation in the presence of contiguous cues. This step is important as a condition needed to recognize the scope depending on the number of cues that the sentence contains.

5.4. Scope Recognition

The scope recognition task extracts the tokens affected by a cue. When negation or speculation is detected using a Prefix-cue, the scope is to the right of that cue (forward in the sentence). On the other hand, if it was detected with a Postfix-cue, the scope is to the left of the cue (backward in the sentence).

This step receives three elements for recognizing the scope: the sentence, the cue, and the regular expression used in the cue detection task. The sentence has to be previously tokenized, and POS (*Part of speech)* tagged. The scope is extracted using five rules: the first four rules were adapted from [57] proposal in order to recognize the scope for both speculation and negation cues. Additionally, a fifth rule that uses syntactic parse trees for scope recognition was added. These rules are detailed as follows:

- *Rule 1:* if the sentence contains a termination term, the scope is extracted using this term. A termination term is a word that indicates the end of the scope. Termination terms are previously created in a lexicon. In the next example the word *"pero"* (but) indicates the end of the scope (the scope is underlined).

  - ***"Probable*** *carcinoma de pulmón,* ***pero*** *espero resultados de biopsia."*
  ("Probable lung carcinoma, but I will wait for biopsy results.")

- *Rule 2:* if a cue C_1 is detected in a sentence containing contiguous cues $C_1, C_2, C_3, \ldots, C_n$, the scope for each C_i will be given by the position of C_{i+1}. For example in the sentence:

  - *"**No** dolor, **no** inflamación, **no** dolores articulares, **ni** fiebre.* (No pain, no inflammation, no joint pain, no fever.), for each cue, the scope is given by the position of the next cue.

- *Rule 3:* if the sentence length corresponds to *"a short sentence heuristic"*, then the scope is given by the end of the sentence. The following example shows a short sentence with their cue and the scope (underlined):

- "***Posible*** *cancer pulmonar*." ("Possible lung cancer.")

- *Rule 4:* if the sentence contains a token POS tagged with a conjunction or verb category. In this case, the scope is determined by the position of this token. The next sentence contains the token *"proponemos"*, which is tagged with the *Verb* category and indicates the end of the scope (the scope is underlined).

  - "Con la ***sospecha*** *de cancer de pulmón, proponemos reunión del comité de tumores."*
  (Given the possibility of lung cancer, we propose a meeting with the tumor committee.)

- *Rule 5:* if the sentence does not match the previous rules, the algorithm generates a sentence parse tree. In this case, the scope is given by the sub-tree that contains the uncertainty or negation cue, as is shown in the next sentence. (The scope is underlined)

  - "Signos ***sugestivos*** *de nódulos pulmonares, con fiebre alta desde ayer."*
  (Signs that suggest pulmonary nodules, with a high fever present since yesterday.)

Figure 6 shows the parse tree generated for the above sentence. In this case, the scope is extracted from the sub-tree where the cue *"sugestivos"* is located. We use UDPipe [59], an open-source NLP tool that performs a dependency parsing tree. The fifth rule is useful in sentences that do not contain any token that indicates the end of the scope.

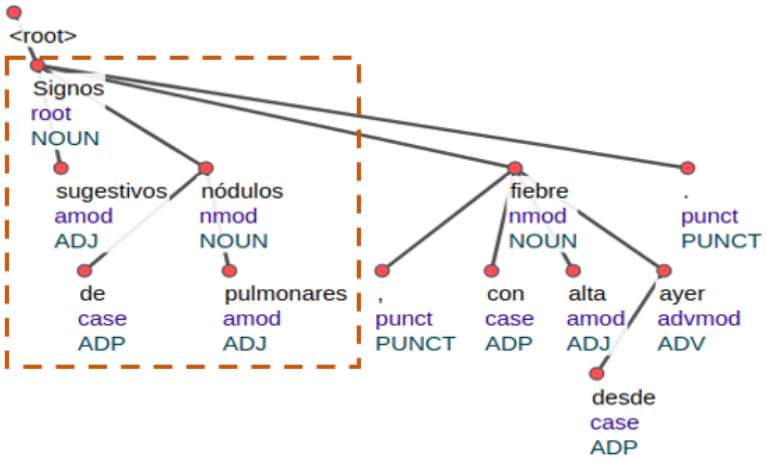

Figure 6. Scope using a Parse tree.

6. Relating Cancer Diagnosis and Dates

Once lung cancer named entities have been extracted and negation and speculation detection has been solved, the only task that rests to extract the cancer diagnosis is relating cancer entities to dates. This task is performed in two steps: (i) linking dates to cancer entities, and (ii) choosing from all the previous linkages the proper diagnosis and its date.

6.1. Linking Dates to Cancer Entities

Cancer entities and dates extracted in a previous step have to be correctly linked. The challenge is how to relate a certain cancer entity to a particular date appearing in the same sentence. To achieve this task, we propose to analyze the syntactic structure of the sentence and the entity events that appear in this sentence. Thus the syntactic parse tree of the sentence is generated, and traversing this tree; it is possible to relate the date to the cancer entity. In particular, a date is linked to a cancer entity if some of the following conditions are true:

1. A cancer entity or an event entity is an ancestor of a date in the dependency path of the sentence.
2. The cancer entity and a date entity appear contiguously or belong to the same predicate in the sentence.

3. The sentence contains event entities: in this case, the event, the cancer entity, and the date are linked if some of the prior conditions are met.

Figure 7 depicts the process to link a cancer entity to a date entity for the sentence: *"Paciente diagnosticado con cáncer de pulmón el 12 October 2016, tratado con cirugía el 21 March 2017"*. (Patient diagnosed with lung cancer on 12 October 2016, treated with surgery on 21 March 2017).

As one can see, the date *"12 October 2016"* is linked to the concept *"cancer de pulmón"* because they are in the same predicate. Moreover, the event *"diagnosticado"* is also linked to this date. After linking these concepts, a triplet *("cancer de pulmón, 12 October 2016, diagnosticado")* is created. On the other hand, the date *"21 March 2017"* is linked to the event *"tratado"* because this event is an ancestor of this date.

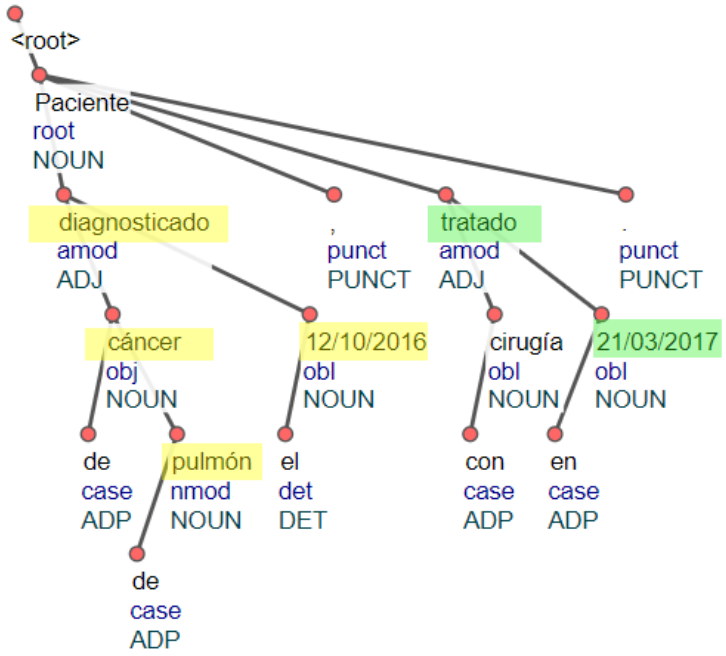

Figure 7. Parse tree: linking dates and cancer diagnosis.

6.2. Choosing the Proper Linkage

The next challenge to be solved after linking dates to cancer entities is to choose the appropriate linkage. As for the same patient, different documents are being annotated from different moments in time; the main challenge is to decide which events refer to diagnosis, and to then eliminate those linkages in which the diagnosis is not clear. This is a challenge because for each patient, hundreds of annotations containing cancer concepts and dates can be extracted. Table 2 shows a possible set of records extracted for the same patient in which we can see instances of triplets: *cancer diagnosis concept, date and event*. Observing the instances one can see that:

- There are different concepts indicating cancer diagnosis entity. Some of them are very generic ("Cancer"), while others are specific terms ("Small cell lung carcinoma").
- There are different dates associated to cancer diagnosis entities.
- Not every record contains an event entity.

Consequently, among all the records associated with a patient, some heuristics are required to find the proper record that matches the diagnosis date. To solve this problem, we propose: (i) to choose the most specific diagnosis, (ii) if more than one record exists associated to this diagnose, but with different dates, then we need to discriminate the most proper date.

Table 2. A set of extracted records.

Cancer Diagnosis Concept	Date	Event
Lung cancer	25 March 2014	
Cancer	March 2015	Begins treatment
Adenocarcinoma	17 May 2017	
Lung carcinoma	14 January 2014	Begins clinical trial
Small cell lung carcinoma	July 2014	Diagnosed
Lung neoplasm	12 July 2016	
Cancer	17 May 2014	
Adenoca	April 2018	
Carcinoma	18 September 2017	Begins surgery

- *Choosing the cancer diagnosis:* A ranked list of UMLS identifiers is used. This list contains UMLS codes and their respective cancer diagnosis sorted according to those that more specifically describe the diagnosis. In this list, the concept *"Squamous cell lung carcinoma"* is more relevant than the concept "Lung cancer" because the former describes more specifically the patient's diagnosis.
- *Choosing the diagnosis date:* two heuristics are applied to disambiguate the date for the chosen diagnosis:
  - Annotations containing events other than *"diagnosed event"* are eliminated.
  - Clinical notes are first ordered chronologically and then classified according to their type: (*Anamnesis, Clinical Judgment, Medical Evolution, treatment, etc.*). According to this classification, the date is assigned taking into account the earliest annotation coming from a document classified as Clinical Judgement or alike.

7. Validation and Results

As explained, the proposed approach is composed of three steps: (i) named entity recognition; (ii) negation and speculation detection; and (iii) linking dates to cancer entities.The validation methodology that was used will be presented first, followed by the results that were obtained for each step. For validation purposes, the following standard measurements were used: Precision (P), Recall (R), and F-score (F1). An implementation for the proposed approach can be found in GitHub (https://github.com/solarte7/lung_cancer_diagnosis).

7.1. Validation Methodology

- To evaluate the deep learning model (Figure 3) the corpus described in Section 4.1 was used, this corpus contains 14,750 annotated sentences. The corpus was shuffled and randomly split into three sets: training (80%), development (10%), and test (10%). This procedure was independently repeated ten times, while verifying that each set contained all the labels in the corpus. The test set was used to calculate the performance metrics, as shown in the Equations (1) and (2). The F-score (Equation (3)) is calculated as a weighted average of the Precision and Recall measurements.
  The BiLSTM-CRF model was developed using TensorFlow (https://www.tensorflow.org/?hl=es-419) and Keras (https://keras.io/) using the following parameters: learning rate as 0.001, dropout as 0.5, the number of epochs is 30, the BiLSTM hidden size is 300, and the batch size is 512.

$$\textbf{Precision} = \frac{\text{Number of entities correctly predicted}}{\text{Number of predicted entities}} \quad (1)$$

$$\textbf{Recall} = \frac{\text{Number of entities correctly predicted}}{\text{Number of entities in the test set}} \quad (2)$$

$$\textbf{F-score} = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}} \quad (3)$$

- To validate the negation and speculation detection step, the NUBES corpus proposed by [12] was used. This public corpus is annotated with speculation and negation in clinical notes written in Spanish. We analyze the performance for each sub-task: *cue detection* and *scope recognition*.
  - A speculation or negation cue is correctly detected when given a sentence, the rule-based approach is able to recognize the cues indicated by the test corpus.
  - A scope is correctly detected when given a sentence and a detected cue, the approach is able to recognize the scope indicated in the corpus.

  Equations (4) and (5) are used to calculate Precision and Recall for the cue detection sub-task. On the other hand, Equations (6) and (7) are used to calculate Precision and Recall for the scope detection sub-task. The F-score is calculated as a weighted average of the Precision and Recall.

$$\textbf{Cue Precision} = \frac{\text{Number of cues correctly detected}}{\text{Number of detected cues}} \quad (4)$$

$$\textbf{Cue Recall} = \frac{\text{Number of cues correctly detected}}{\text{Number of the cues in the corpus}} \quad (5)$$

$$\textbf{Scope Precision} = \frac{\text{Number of scopes correctly detected}}{\text{Number of detected scopes}} \quad (6)$$

$$\textbf{Scope Recall} = \frac{\text{Number of scopes correctly detected}}{\text{Number of the scopes in the corpus}} \quad (7)$$

- To validate the diagnosis date extraction, a database containing data from *"Hospital Universitario Puerta de Hierro Madrid"* was used. It contains around 300,000 clinical notes corresponding to 1000 patients that were diagnosed with lung cancer in the last ten years. Information of the diagnosis date for each patient is available. A diagnosis date is correctly extracted when it corresponds to the diagnosis date given by the hospital dataset. Equations (8) and (9) are used to calculate the Precision and Recall respectively.

$$\textbf{Date Precision} = \frac{\text{Number of diagnosis date correctly extracted}}{\text{Number of diagnosis date extracted}} \quad (8)$$

$$\textbf{Date Recall} = \frac{\text{Number of diagnosis date correctly extracted}}{\text{Number of diagnosis date in the hospital dataset}} \quad (9)$$

7.2. Named Entity Recognition Results

For each split in the corpus (training, development, test) we carry out the following experiments:

1. The BiLSTM-CRF base model proposed by [32] is used.
2. General domain embeddings training by Fast text (https://fasttext.cc/docs/en/pretrained-vectors.html) on Wikipedia are added to the BiLSTM-CRF model.
3. Spanish medical embeddings proposed by [41] have been added to the BiLSTM-CRF model.
4. Spanish medical embeddings and char embeddings are added to the model.
5. A combination of medical embeddings, char embeddings, lemmas, and Part of Speech (POS) tagging features.

Table 3 shows obtained results for every experiment previously described, and the partition (training, development, test) that obtained the best performance was chosen. According to this table, one can see that the best results are obtained in the fifth experiment. This fact indicates that the combination of medical embeddings, char embeddings, lemmas,

and POS tagging features considerably improve cancer entity extraction. Moreover, the use of medical embeddings has the most significant impact on improving the model. This fact suggests that medical domain embeddings helps the neural network to learn to extract named entities. In addition, the use of lemmas and POS tagging features also have a considerable impact on the model, while adding char embeddings seems to have less impact on the final rate.

On the other hand, (see Table 3), the use of Wikipedia general domain embeddings does not obtain such effective results. This fact suggests that using general domain embeddings to extract entities in specific and specialized domains such as the medical domain does not produce the best results.

The proposed model's performance is similar to other proposals dealing with extracting cancer entities from clinical notes written in Spanish [45–47]. However, the main limitation with those proposals is that they are only able to extract cancer entities and do not consider other entities required to correctly extract the cancer diagnosis, such as dates, events, family members, or tumor stage.

Table 3. Results for Named Entity Recognition.

Model	P	R	F1
BiLSTM-CRF	0.83	0.78	0.80
BiLSTM-CRF + General domain embeddings	0.80	0.73	0.76
BiLSTM-CRF + Medical embeddings	0.87	0.84	0.85
BiLSTM-CRF + Medical Embeddings + Char embeddings	0.87	0.85	0.86
BiLSTM-CRF + Medical Embeddings + Char embeddings + Lemmas + POS	**0.91**	**0.89**	**0.90**

7.3. Negation and Speculation Results

7.3.1. Cue Detection Results

Three experiments were performed to measure the impact of regular expressions (Regex) to detect both negation and speculation cues:

1. Only the first and second Regex are used for cue detection (See Section 5.2). These Regex were adapted to Spanish from the popular rule-based Negex [56] proposal. The Negex adaptation to Spanish is used as baseline.
2. The third Regex is added to measure the contiguous cues behavior.
3. The fourth Regex was added, this experiment includes all Regex proposed in Section 5.2.

Table 4 shows the results obtained in the cue detection task. The best performance using all the Regex proposed in Section 5.2, for both negation and speculation cue detection was obtained. In the case of negation detection, adding the third Regex has a significant improvement in the F-score, as evidenced in Table 4. This fact suggests that detecting contiguous cues improves negation detection in clinical notes written in Spanish. Adding the fourth Regex also improves performance but to a lesser improvement rate.

On the other hand, for speculation detection, using the first two Regex an F-score of 89% is obtained. When adding the third and fourth Regex, the improvement rate is not as significant as in negation detection. This fact suggests that contiguous cues do not have an important impact on speculation detection.

Table 4. Results for the cue detection task.

	Negation			Speculation		
Regex	**P**	**R**	**F1**	**P**	**R**	**F1**
Regex 1,2	0.84	0.81	0.82	0.91	0.88	0.89
Regex 1,2,3	0.93	0.91	0.92	0.92	0.89	0.90
Regex 1,2,3,4	**0.96**	**0.94**	**0.95**	**0.92**	**0.89**	**0.90**

7.3.2. Scope Recognition Results

Five experiments were performed in order to analyze the impact for each rule proposed for scope recognition (see Section 5.4):

1. The first rule is used to recognize the scope. As mentioned in Section 5.4, this rule searches for a termination term in the sentence which indicates the end of the scope. This approach is commonly used by previous rule-based approaches [53,60,61].
2. The second rule is added to take into account contiguous cues for recognizing the scope.
3. The *short sentence heuristic* is used by adding the third rule.
4. Adding the fourth rule to include POS tagging features.
5. The fifth rule is added to include the parse tree analysis for extracting the scope.

Table 5 shows the results obtained in the scope recognition task. According to this table, the best results were obtained when all the proposed rules were included in the fifth experiment. An 89% F-score in the negation scope, and an 85% F-score in speculation scope detection were obtained. This fact indicates that the proposed approach outperforms previous rule-based approaches that use only termination terms for scope recognition in Spanish [60,61]. In fact, when using only termination terms, a mere 70% in F-score for negation and an 71% in F-score for speculation detection were obtained. This is due to the fact that not all sentences have termination terms.

For the particular case of negation detection, Table 5 depicts that the use of contiguous negation, short sentence heuristic, POS tagging, and sentence parse tree (Rules 2 to 5) shows to be useful to improve the scope recognition task. All these rules have similar improvement rates. This proposal improves previous rule-based approaches which recognize the scope using only termination terms [60,61].

In speculation scope detection, the behavior of the rules were found to be different. Specifically, the contiguous cues and the *short sentence heuristic* do not have a significant improvement rate (See Table 5). This situation suggests that contiguous cues and short sentences are not so related to speculation as they are for negation. On the contrary, POS tagging and sentence parse tree features result in a significant improvement for speculation scope detection.

In addition, speculation detection has not yet been fully addressed for Spanish clinical narratives. We found only the [12] proposal dealing with this issue. Obtained results in our approach show similar performance rates than those reported by [12].

Table 5. Results for the scope recognition task.

	Negation			Speculation		
Rules	**P**	**R**	**F1**	**P**	**R**	**F1**
Rule 1	0.72	0.70	0.71	0.72	0.69	0.70
Rules 1,2	0.78	0.75	0.76	0.72	0.69	0.70
Rules 1, 2, 3	0.81	0.79	0.80	0.73	0.71	0.72
Rules 1, 2, 3, 4	0.86	0.82	0.84	0.82	0.79	0.80
Rules 1, 2, 3, 4, 5	**0.91**	**0.87**	**0.89**	**0.87**	**0.84**	**0.85**

7.4. Relating Cancer Diagnosis and Date Results

The impact of negation and speculation to extract correctly the diagnosis date will be evaluated as it is shown in the planned experiments:

1. The diagnosis date is chosen from all extracted named entities which belong to the patient.
2. The diagnosis date is chosen after filtering entities affected by negation.
3. The diagnosis date is chosen after filtering entities affected by negation and speculation.

Table 6 shows obtained results for linking the proper date to the lung cancer diagnosis. According to this table, the best result was obtained in the third experiment with an 89% F-score. This fact suggests that detecting negation and speculation is a crucial step needed to extract a diagnosis date correctly. In fact, if the diagnosis date is extracted using only results from named entity recognition, a 64% F-score is obtained.

In addition, according to Table 6, filtering entities affected by negation improves the F-score by 7%. Meanwhile, filtering entities affected by speculation improves the F-score by 18%. This indicates that the diagnosis date extraction is more sensitive to speculation than to negation. Moreover, when comparing the results of the proposed approach to those presented in [10], the F-score performance rate is improved by 5%, which is mainly due to the improving of cancer concept extraction and speculation detection.

Table 6. Results for extracting the diagnosis date.

Experiment	P	R	F1
Using all Named Entities	0.67	0.62	0.64
Filtering negated entities	0.72	0.71	0.71
Filtering negated and speculated entities	**0.92**	**0.87**	**0.89**

8. Discussion

Automatically extracting lung cancer diagnosis from clinical notes can be useful to support clinical research as establishing the diagnosis is crucial to understand factors occurring prior and after the disease. The approach proposed in this paper contributes to automatically extract diagnosis and dates from clinical narratives written in Spanish.

The use of deep learning methods was effective for named entity recognition in the cancer domain. The performance of the BiLSTM-CRF model has shown promising results as it can see in Table 3. Specifically, including the Event entity was useful to disambiguate sentences containing several date mentions. Thus our approach outperforms the proposal presented in [10] as the latter does not take into account event extraction. Besides, it is important to note that once a neural network has been trained with lung cancer-related data it can be used to learn not only on lung cancer notes but also on other kind of tumors.

On the other hand, the rule-based approach presented has shown to return competitive results for speculation and negation detection tasks. In particular, the following facts should be stressed:

- The proposed approach uses a lexicon that has been adapted to Spanish from two different resources specialized in negation and speculation in the biomedical domain. Additionally, the lexicon was semantically extended with a word embeddings technique that helped obtain new cues with similar meanings. Moreover, this lexicon was manually reviewed and evaluated. The resulting lexicon offers high precision in the cue detection task.
- Some weaknesses reported in [12], such as *post-scope recognition* are addressed in this proposal by using Regex 2 (see Section 5.2). When a cue is detected using this Regex, the scope is searched to the left of the cue.
- The rules proposed for scope recognition are adapted according to the analysis that has been performed on how speculation and negation are expressed in clinical notes written in Spanish (see Section 5.3).

Despite the promising results, still some limitations have to be addressed. In particular, one of the main challenges for using deep learning approaches is that one must have an annotated corpus. In this paper, a manually annotated corpus to support lung cancer diagnosis extraction was proposed. However, a larger corpus containing more annotations should be addressed in future experiments. Moreover, to extract more information regarding lung cancer disease, a more comprehensive set of entities would be needed.

In the speculation cue detection, one of the challenges to be solved relates to the detecting of syntactic speculation cues, specifically the token *"o"* (or). This token does

not behave as a speculation cue in more than 99% of cases. In order to illustrate this, consider the following sentences:

1. *"Causas de sangrado: una complicación de la cirugía **o** una patología asociada al cancer de pulmón."* (Causes of bleeding: a surgery-related complication **or** a lung cancer-related pathology)
2. *"Si acude a alguna consulta **o** servicio sanitario debe ser acompañado. "* (Patients must be accompanied when going to general appointments **or** when using the health service.))

In the first sentence, the token *"o"* acts as a speculation cue, while in the second one, it acts as a disjunction element. As the second case is the most frequent, in our approach we decided not to include the token *"o"* in the cue lexicon (Figure 4) in order to reduce false positives detection. However this should be improved in a future research .

In the scope recognition task, the main limitation of the presented approach is related to discontinuous scopes detection. Although this situation is infrequent, it should be solved. In the NUBES corpus [12] discontinuous scope represents less than 5% of the cases. Two cases where a discontinuous scope can appear were identified:

- Sentences containing a sequence of contiguous cues where some part of the text is not affected by any cue. In the next sentence the scope is underlined and the text *"sangrado pulmonar"* is out the scope.

  *"**No** vómitos, sangrado pulmonar desde ayer, **no** tos, **no** dolor."*

  (No vomiting, lung bleeding since yesterday, no cough, no pain.)
- When the scope of a cue is in both directions, to the left and to the right of the cue. In the next sentence, the cue is the word "Versus", and the scope is underlined.

  Cancer de pulmón ***Versus*** *Infección en lóbulo derecho.*

  (Lung cancer Versus right lobe infection).

These cases should be further analyzed in the future. Another challenge can be found in the fact that the variability of speculation cues is higher than when dealing with negation cues. This fact increases the complexity of the speculation detection task.

9. Conclusions and Future Work

Automatically extracting lung cancer diagnosis and its diagnosis date from clinical narratives written in Spanish has been approached in this paper through a process divided into three steps: Named entity recognition (NER), negation and speculation detection, and linking the proper date to the cancer diagnosis. This approach improves the lung cancer diagnosis extraction process by properly relating a cancer concept to the diagnosis date.

Deep learning-based approaches have shown to be useful to improve named entity recognition in the medical domain. Specifically, a BiLSTM-CRF model to extract lung cancer entities from clinical narratives written in Spanish was implemented, and the results that were obtained are encouraging.

Speculation and negation detection is a crucial step to improve information extraction in the medical domain. In particular, speculation detection highly impacts the accuracy of diagnosis extraction. For this reason, properly filtering speculative cancer concepts is an important factor for correctly extracting the diagnosis date.

Extracting useful information from clinical notes, and in particular, accurate cancer information, is a promising task to improve clinical decision support systems. The ability to analyze clinical texts written in Spanish opens important opportunities to develop more clinical applications. The presented approach contributes to this line of research. Approaches to improve negation and speculation detection and named entity recognition will be explored in future studies.

Author Contributions: O.S.P. has been responsible of the design and implementation of the proposed solution. He has run the experiments and he is responsible of the manuscript writing. M.T. and M.P. as clinicians are responsible of medical validation of results and responsible of setting the problem and helping on finding the concepts to be found. E.M. and A.R.-G. have been responsible of

the technical coordination of the work, they have also coordinate writing of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the EU Horizon 2020 innovation program under grant agreement No. 780495, project BigMedilytics (Big Data for Medical Analytics). It has been also supported by Fundación AECC and Instituto de Salud Carlos III (grant AC19/00034), under the frame of ERA-NET PerMed.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spasić, I.; Livsey, J.; Keane, J.A.; Nenadić, G. Text mining of cancer-related information: Review of current status and future directions. *Int. J. Med. Inform.* **2014**, *83*, 605–623. [CrossRef] [PubMed]
2. Demner-Fushman, D.; Chapman, W.W.; McDonald, C.J. What can natural language processing do for clinical decision support? *J. Biomed. Inform.* **2009**, *42*, 760–772. [CrossRef]
3. Zeng, Q.T.; Goryachev, S.; Weiss, S.; Sordo, M.; Murphy, S.N.; Lazarus, R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.* **2006**, *30*, 327–348. [CrossRef]
4. Wang, L.; Luo, L.; Wang, Y.; Wampfler, J.; Yang, P.; Liu, H. Natural language processing for populating lung cancer clinical research data. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1–10. [CrossRef] [PubMed]
5. Zhang, X.; Zhang, Y.; Zhang, Q.; Ren, Y.; Qiu, T.; Ma, J.; Sun, Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int. J. Med. Inform.* **2019**, *132*, 103985. [CrossRef]
6. Sheikhalishahi, S.; Miotto, R.; Dudley, J.T.; Lavelli, A.; Rinaldi, F.; Osmani, V. Natural language processing of clinical notes on chronic diseases: Systematic review. *J. Med. Internet Res.* **2019**, *21*, 1–18. [CrossRef] [PubMed]
7. de Groot, P.M.; Wu, C.C.; Carter, B.W.; Munden, R.F. The epidemiology of lung cancer. *Transl. Lung Cancer Res.* **2018**, *7*, 220–233. [CrossRef] [PubMed]
8. Najafabadipour, M.; Tuñas, J.M.; Rodríguez-González, A.; Menasalvas, E. Lung Cancer Concept Annotation from Spanish Clinical Narratives. In *Data Integration in the Life Sciences*; Auer, S., Vidal, M.E., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 153–163.
9. Savova, G.K.; Tseytlin, E.; Finan, S.P.; Castine, M.; Timothy, A.; Medvedeva, O.P.; Harris, D.A.; Hochheiser, H.S.; Lin, C.; Girish, R. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. *Cancer Res.* **2017**, *77*, 1–6. [CrossRef]
10. Solarte-Pabon, O.; Torrente, M.; Rodriguez-Gonzalez, A.; Provencio, M.; Menasalvas, E.; Tunas, J.M. Lung cancer diagnosis extraction from clinical notes written in spanish. In Proceedings of the IEEE Symposium on Computer-Based Medical Systems, Rochester, MN, USA, 28–30 July 2020; pp. 492–497. [CrossRef]
11. Alam, R.; Cheraghi-Sohi, S.; Panagiot, M.; Esmail, A.; Campbell, S.; Panagopoulou, E. Managing diagnostic uncertainty in primary care: A systematic critical review. *BMC Fam. Pract.* **2017**, *18*, 1–13. [CrossRef] [PubMed]
12. Lima, S.; Perez, N.; Cuadros, M.; Rigau, G. NUBES: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. *arXiv* **2020**, arXiv:2004.01092.
13. Cruz Díaz, N.P.; Maña López, M.J. *Negation and Speculation Detection*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2019; [CrossRef]
14. Agarwal, S.; Yu, H. Detecting hedge cues and their scope in biomedical text with conditional random fields. *J. Biomed. Inform.* **2010**, *43*, 953–961. [CrossRef] [PubMed]
15. Fu, S.; Chen, D.; He, H.; Liu, S.; Moon, S.; Peterson, K.J.; Shen, F.; Wang, L.; Wang, Y.; Wen, A.; et al. Clinical concept extraction: A methodology review. *J. Biomed. Inform.* **2020**, *109*, 103526. [CrossRef]
16. Tulkens, S.; Šuster, S.; Daelemans, W. Unsupervised concept extraction from clinical text through semantic composition. *J. Biomed. Inform.* **2019**, *91*, 103120. [CrossRef] [PubMed]
17. Yim, W.W.; Yetisgen, M.; Harris, W.P.; Sharon, W.K. Natural Language Processing in Oncology Review. *JAMA Oncol.* **2016**, *2*, 797–804. [CrossRef]
18. Warner, J.L.; Levy, M.A.; Neuss, M.N.; Warner, J.L.; Levy, M.A.; Neuss, M.N. ReCAP: Feasibility and Accuracy of Extracting Cancer Stage Information from Narrative Electronic Health Record Data. *J. Oncol. Pract.* **2016**, *12*, 157–158. [CrossRef]
19. Nguyen, A.N.; Lawley, M.J.; Hansen, D.P.; Bowman, R.V.; Clarke, B.E.; Duhig, E.E.; Colquist, S. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 440–445. [CrossRef] [PubMed]
20. Deshmukh, P.R.; Phalnikar, R. TNM Cancer Stage Detection from Unstructured Pathology Reports of Breast Cancer Patients. In Proceedings of the International Conference on Computational Science and Applications, Pune, India, 7–9 August 2019; Bhalla, S., Kwan, P., Bedekar, M., Phalnikar, R., Sirsikar, S., Eds.; pp. 411–418.
21. AAlAbdulsalam, A.K.; Garvin, J.H.; Redd, A.; Carter, M.E.; Sweeny, C.; Meystre, S.M. Automated Extraction and Classification of Cancer Stage Mentions from Unstructured Text Fields in a Central Cancer Registry. *AMIA Jt. Summits Transl. Sci. Proc.* **2018**, *2017*, 16–25.

22. Evans, T.L.; Gabriel, P.E.; Shulman, L.N. Cancer Staging in Electronic Health Records: Strategies to Improve Documentation of These Critical Data. *J. Oncol. Pract.* **2016**, *12*, 137–139. [CrossRef] [PubMed]
23. Khor, R.C.; Nguyen, A.; O'Dwyer, J.; Kothari, G.; Sia, J.; Chang, D.; Ng, S.P.; Duchesne, G.M.; Foroudi, F. Extracting tumour prognostic factors from a diverse electronic record dataset in genito-urinary oncology. *Int. J. Med. Inform.* **2019**, *121*, 53–57. [CrossRef]
24. Wang, Z.; Shah, A.D.; Tate, A.R.; Denaxas, S.; Shawe-Taylor, J.; Hemingway, H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS ONE* **2012**, *7*, [CrossRef]
25. Zheng, S.; Jabbour, S.K.; O'Reilly, S.E.; Lu, J.J.; Dong, L.; Ding, L.; Xiao, Y.; Yue, N.; Wang, F.; Zou, W. Automated Information Extraction on Treatment and Prognosis for Non–Small Cell Lung Cancer Radiotherapy Patients: Clinical Study. *JMIR Med. Inform.* **2018**, *6*, e8. [CrossRef]
26. Bitterman, D.; Miller, T.; Harris, D.; Lin, C.; Finan, S.; Warner, J.; Mak, R.; Savova, G. Extracting Relations between Radiotherapy Treatment Details. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, Online Conference, 19 November 2020; pp. 194–200. [CrossRef]
27. Zeng, Z.; Espino, S.; Roy, A.; Li, X.; Khan, S.A.; Clare, S.E.; Jiang, X.; Neapolitan, R.; Luo, Y. Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinform.* **2018**, *19*. [CrossRef]
28. Isaksson, L.J.; Pepa, M.; Zaffaroni, M.; Marvaso, G.; Alterio, D.; Volpe, S.; Corrao, G.; Augugliaro, M.; Starzyńska, A.; Leonardi, M.C.; et al. Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy. *Front. Oncol.* **2020**, *10*. [CrossRef]
29. Forsyth, A.W.; Barzilay, R.; Hughes, K.S.; Lui, D.; Lorenz, K.A.; Enzinger, A.; Tulsky, J.A.; Lindvall, C. Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms From Electronic Health Records. *J. Pain Symptom Manag.* **2018**, *55*, 1492–1499. [CrossRef]
30. Hochreiter, S.; Schmidhuber, J. LSTM can solve hard long time lag problems. In Proceedings of the 9th International Conference on Neural Information Processing Systems, Denver, CO, USA, 2–5 December 1996; pp. 473–479.
31. Goldberg, Y. Neural Network Methods for Natural Language Processing. *Synth. Lect. Hum. Lang. Technol.* **2017**, *10*, 1–311. [CrossRef]
32. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016—Proceedings of the Conference, San Diego, CA, USA, 12–17 June 2016; pp. 260–270. [CrossRef]
33. Lopez, M.M.; Kalita, J. Deep Learning applied to NLP. *arXiv* **2017**, arXiv:1703.03091.
34. Carta, S.; Ferreira, A.; Podda, A.S.; Reforgiato Recupero, D.; Sanna, A. Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting. *Expert Syst. Appl.* **2021**, *164*, 113820. [CrossRef]
35. Vázquez, J.J.; Arjona, J.; Linares, M.; Casanovas-Garcia, J. A Comparison of Deep Learning Methods for Urban Traffic Forecasting using Floating Car Data. *Transp. Res. Procedia* **2020**, *47*, 195–202. [CrossRef]
36. Nguyen, G.; Dlugolinsky, S.; Tran, V.; Lopez Garcia, A. Deep learning for proactive network monitoring and security protection. *IEEE Access* **2020**, *8*, 19696–19716. [CrossRef]
37. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Nice, France, 2013; pp. 3111–3119.
38. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
39. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
40. Wang, Y.; Liu, S.; Afzal, N.; Rastegar-Mojarad, M.; Wang, L.; Shen, F.; Kingsbury, P.; Liu, H. A comparison of word embeddings for the biomedical natural language processing. *J. Biomed. Inform.* **2018**, *87*, 12–20. [CrossRef]
41. Soares, F.; Villegas, M.; Gonzalez-Agirre, A.; Krallinger, M.; Armengol-Estapé, J. Medical word embeddings for Spanish: Development and evaluation. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 124–133. [CrossRef]
42. Névéol, A.; Dalianis, H.; Velupillai, S.; Savova, G.; Zweigenbaum, P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *J. Biomed. Semant.* **2018**, *9*, 1–13. [CrossRef]
43. Najafabadipour, M.; Zanin, M.; Rodriguez-Gonzalez, A.; Gonzalo-Martin, C.; Garcia, B.N.; Calvo, V.; Bermudez, J.L.C.; Provencio, M.; Menasalvas, E. Recognition of time expressions in Spanish electronic health records. In Proceedings of the IEEE Symposium on Computer-Based Medical Systems, Cordoba, Spain, 5–7 June 2019; pp. 69–74. [CrossRef]
44. Wang, L.; Wampfler, J.; Dispenzieri, A.; Xu, H.; Yang, P.; Liu, H. Achievability to Extract Specific Date Information for Cancer Research. In *AMIA Annual Symposium Proceedings, AMIA Symposium*; American Medical Informatics Association: Washington, DC, USA, 2019; Volume 2019, pp. 893–902.
45. Garciá-Pablos, A.; Perez, N.; Cuadros, M. Vicomtech at cantemist 2020. *CEUR Workshop Proc.* **2020**, *2664*, 489–498.

46. Carrasco, S.S.; Martínez, P. Using embeddings and bi-lstm+crf model to detect tumor morphology entities in Spanish clinical cases. *CEUR Workshop Proc.* **2020**, *2664*, 368–375.
47. López-Úbeda, P.; Diáz-Galiano, M.C.; Martín-Valdivia, M.T.; Urenã-López, L.A. Extracting neoplasms morphology mentions in Spanish clinical cases throughword embeddings. *CEUR Workshop Proc.* **2020**, *2664*, 324–334.
48. Miranda-Escalada, A.; Farré, E.; Krallinger, M. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results. *CEUR Workshop Proc.* **2020**, *2664*, 303–323.
49. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Williamstown, MA, USA, 28 June–1 July 2001; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 282–289.
50. De Albornoz, J.C.; Plaza, L.; Diaz, A.; Ballesteros, M. UCM-I: A rule-based syntactic approach for resolving the scope of negation. In Proceedings of the *SEM 2012—1st Joint Conference on Lexical and Computational Semantics, Montréal, QC, Canada, 7–8 June 2012; Volume 1, pp. 282–287.
51. Dalianis, H. *Clinical Text Mining*; Springer: Berlin/Heidelberg, Germany, 2018; [CrossRef]
52. Vincze, V.; Szarvas, G.; Farkas, R.; Móra, G.; Csirik, J. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinform.* **2008**, *9*, 1–9. [CrossRef] [PubMed]
53. Harkema, H.; Dowling, J.N.; Thornblade, T.; Chapman, W.W. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. Biomed. Inform.* **2009**, *42*, 839–851. [CrossRef]
54. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
55. Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
56. Chapman, W.W.; Bridewell, W.; Hanbury, P.; Cooper, G.F.; Buchanan, B.G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.* **2001**, *34*, 301–310. [CrossRef]
57. Solarte-Pabón, O.; Menasalvas, E.; Rodriguez-González, A. Spa-neg: An approach for negation detection in clinical text written in Spanish. In Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain, 6–8 May 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 323–337. [CrossRef]
58. Elazhary, H. NegMiner: An automated tool for mining negations from electronic narrative medical documents. *Int. J. Intell. Syst. Appl.* **2017**, *9*, 14–22. [CrossRef]
59. Straka, M.; Hajič, J.; Straková, J. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portoroz, Slovenia, 23–28 May 2016; pp. 4290–4297.
60. Stricker, V.; Iacobacci, I.; Cotik, V. Negated Findings Detection in Radiology Reports in Spanish: An Adaptation of NegEx to Spanish. In Proceedings of the Workshop on Replicability and Reproducibility in Natural Language Processing: Adaptive Methods, Resources and Software at IJCAI 2015, Buenos Aires, Argentina, 25–27 July 2015; pp. 1–7.
61. Costumero, R.; Lopez, F.; Gonzalo-Martín, C.; Millan, M.; Menasalvas, E. An approach to detect negation on medical documents in Spanish. In Proceedings of the Brain Informatics and Health, Warsaw, Poland, 11–14 August 2014; pp. 366–375.