Universidad del País Vasco Euskal Herriko Unibertsitatea

End to end approach for i2b2 2012 challenge based on Cross-lingual models

Author: Edgar Andrés Santamaría

Advisors: Aitziber Atutxa Salazar and Oier López de Lacalle

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

September 2020

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Laburpena

Deskribapena Narratiba klinikoen domeinuan i2b2 2012 erronkarako hizkuntzarteko ikuspegia jorratzen duen soluzioa proposatzen dugu. Erronka honek txosten medikuetan islatzen diren gertaeren arteko denbora-erlazioak iragartzea du helburu. Horretarako, lan hau alde batetik (1) klinikoki esanguratsuak diren gertaerak, adibidez, kontzeptu klinikoak, probak, tratamenduak, sail klinikoak eta bestetik, (2) denbora-adierazpenak, adibidez, txostenak esleituta duen data, denbora, iraupen edo maiztasuna adierazten duten espresioak antzeman eta bukatzeko gertaera klinikoen eta (3) denbora-adierazpenen arteako erlazioak anotatuta duen corpus batetik abiatzen da.

Helburuak Lanaren helburuak i2b2 2012 artearen egoera hobetzea eta Cross-lingual modeloa Data baliabide baxuak dituen domeinu kliniko espezifikora egokitzea dira.

Metodoak Lana modulu desberdinetako hobi gisa ulertu da, gertaera eta denbora-adierazpenetarako sekuentzia-markatzaileak, eta denbora-erlaziorako perpaus-sailkatzailea, independenteki garatu dira. XLM-RoBERTa Cross-lingual modeloa erabili izan da lan honetan.

Emaitzak Gertaerak atzemateko, 0.91 Span F1 exekutatzen duen sekuentzia-markatzailea proposatzen dugu. Denbora-adierazpenetarako, 0.91 Span F1 egiten duen sekuentzia-markatzailea bat proposatzen dugu. Denbora-erlaziorako, 0.29 F1 neurria egiten duten sekuentzia-markatzaileetan oinarritutako perpaus-sailkatzailea proposatzen dugu.

Abstract

Background We propose a Cross-lingual approach to i2b2 2012 challenge for Clinical Records focused on the temporal relations in clinical narratives. Corpus of discharge summaries annotated with temporal information was provided for automatically extracting : (1) clinically significant events, including both clinical concepts such as problems, tests, treatments, and clinical departments, and events relevant to the patient's clinical timeline, such as admissions, transfers between departments, etc; (2) temporal expressions, referring to the dates, times, duration, or frequencies in the clinical text. The values of the extracted temporal expressions had to be normalized to an ISO specification standard; and (3) temporal relations, among the clinical events and temporal expressions.

Goals The objectives involved in the current work consists on outperforming previous State of the Art for the i2b2 2012 challenge and adapting Cross-lingual model into clinical specific domain with low Data resources available.

Methods The task has been conceived as a pipeline of different modules, an event and temporal expression token-classifier and a text-classifier for relation extraction, each of them independently developed from the other. We used XLM-RoBERTa Cross-lingual model.

Results For event detection, the proposed token-classifier obtains a 0.91 Span F1. For temporal expressions, our sentence-classifier achieves a 0.91 Span F1. For temporal relation, we propose sentence classifier based on sequential-taggers that performs at 0.29 F1 measure.

Contents

List of Figures

List of Tables

1 Introduction

Nowadays clinical institutions generate huge amounts of clinical histories, those usually compose the historical summary of the patients, and doctors retrieve crucial information in the registers in order to successfully treat the patients taking into account their clinical life. Those files are usually generated while patient is admitted in certain clinical institution, and comprises the relevant information until the discharge. The provided information is mainly based on clinical events which are relevant doings that were performed during the patient stay, for example: treatments, departments, diagnoses, problems, etc. Clinical events are usually related to certain timestamp because Doctors are interested on the timeline progress into heal. The register time stamp comprises wide range of metrics : year, month, day or hour. All the information is provided in written format. The information contained in these records is relevant because describes every clinical process the patient has gone through alongside the timestamp. We aim to provide suitable end-to-end approach for "2012 i2b2 Temporal Relations Challenge" (Sun et al., 2013). the aim is to automatically annotate temporal expressions with relevant medical terms, and establish relations among them as shown in the following figure 1. The system is evaluated using the provided resources of "2012 i2b2 Temporal Relations Challenge".

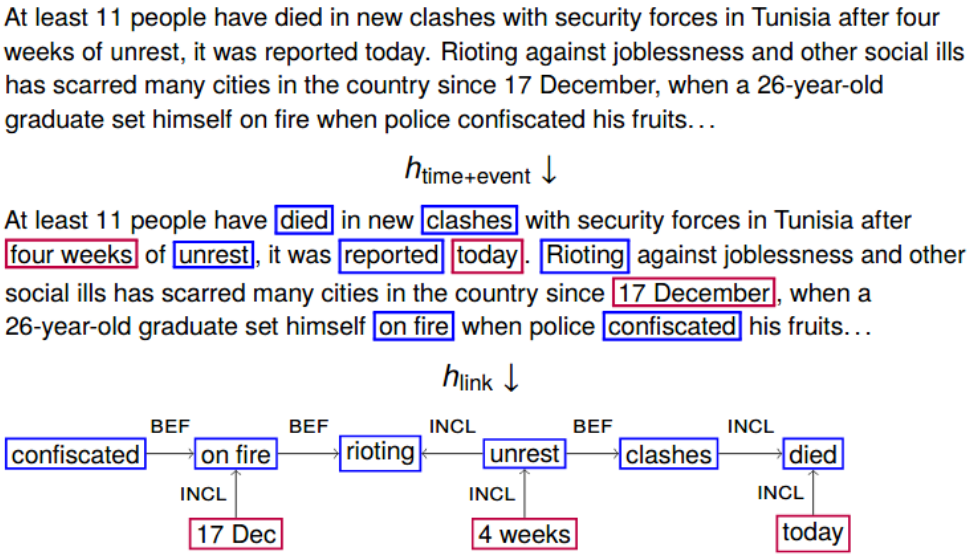

Figure 1: Motivational example extracted from (Steven Bethard, 9)

The objectives involved in the current approach consist on :

1. Implementation of a state-of-the-art system for medical entity and event extraction (EVENT, TIMEX3).
2. Implementation of a state-of-the-art system for end-to-end task.
3. Domain Adaption of the system in low data regime, such as bio-medical domain.

The core technology are transformer based pretrained language models into which we apply transfer learning specifically transductive transfer learning according to (Ruder, 2019) because we have labelled data in English source domain and we can apply the same solution into different languages, even there exist some case of success there isn't sufficient research in domain adaptation for low data regime domains.

It is important to mention language variation as a challenging issue regarding this task. Doctors tend to write in spontaneous language, therefore same term appears in many many different forms because of lack of standardization. The following example shows different forms for the same term:

CT scan was performed.

Computerized Tomography scan was performed.

In addition the intrinsic ambiguity of the language also appears in the task, referred to syntactic even semantic level. In the first case we have to figure if phosphatase was on normal range or she was normal even with phosphatase elevated. In the second we have to figure related to which time we are talking :

Her alkaline phosphatase was slightly elevated but otherwise relatively normal.

She had a normal pancreas at that time.

Finally, task stays within the clinical domain and therefore some specificities regarding the domain like terminology and particular ways to use the common language, as we can see in the following example:

1 cm cyst in the right lobe of the liver.

In this case the metric "1 cm" is referred to the "cyst" diameter, in addition "in the" and "of the" are used to locate "cyst" in the human corpse distribution "right lobe" specifically on "liver". This specific way to use the language based on clinical domain requires language modeling (LM) efforts to aid the models finding patterns over text.

2 Description of i2b2 challenge

In the 2012 i2b2 Challenge (Sun et al., 2013), 310 discharge summaries were annotated for temporal information. The challenge focused specifically on the identification of clinically relevant events in the patient records, and the relative temporal ordering of the events with respect to each other and with respect to time expressions included in the records.

To achieve the goal, the organizers divided the challenge into two main tasks: an event and temporal expressions identification (EVENT and TIMEX3 task) and a relation extraction (TLINK task). For EVENT and TIMEX3 tasks are sequence labelling tasks consisting on annotating all EVENT and TIMEX3 terms in the text by assigning the correct label to each of them. Additionally each EVENT and TIMEX3 has certain attributes like mode, type or value, so it is also necessary to pursue an attribute prediction to provide the correct attributes for the annotated label. TLINK task consists on extracting the relations among the EVENT and TIMEX3 tags. We evaluate the performance using the provided scripts by "i2b2 2012 challenge". Here one example of a discharge summary 2 and its annotation 3 as provided by the organizers.

<ClinicalNarrativeTemporalAnnotation>
<TEXT>
Admission Date : 2012-05-13 Discharge Date : 2012-05-22 Service : Neonatology HISTORY OF PRESENT ILLNESS : Fred Grauman is the former 2.252 kg product of a 35-1/7 week gestatio
pregnancy born to a 37-year-old G4 P3-5 woman . Prenatal screens : Blood type O+ , antibody negative , rubella immune , RPR nonreactive , hepatitis B surface antigen negative , group bet
Strep positive . The pregnancy was complicated by anemia . Mother has a prior history of mitral valve prolapse . She had a prior cesarean section done , two vaginal births after cesarean
section . She presented in labor on the day of delivery . These infants were delivered by repeat cesarean section , this twin emerged with good tone , color , and cry . Apgars were eight at o
minute and nine at five minutes . He was admitted to the Neonatal Intensive Care Unit for treatment of prematurity . HOSPITAL COURSE BY SYSTEMS INCLUDING PERTINENT
LABORATORY DATA: 1. Respiratory : Raymond required some intermittent blow-by oxygen for the first two hours of birth . He has remained in room air since that time . The grunting noted
on admission resolved within the first few hours after birth . He did not have any episodes of spontaneous apnea or bradycardia . 2. Cardiovascular : No murmurs were noted , although
normal heart rates and blood pressures . 3. Fluids , electrolytes , and nutrition : Enteral feeds were started on the date of birth and gradually advanced to full volume . He required some
gavage feeds through day of life #4 , and has been all po since day of life #5 . At the time of discharge , he is taking Enfamil 20 calories / ounce with iron . Discharge weight is 2.37 kg , hea
circumference 32 cm , length 47 cm . 4. Infectious Disease : Due to the unknown group B Strep status and prematurity , Raymond was evaluated for sepsis . A complete blood count had a
white blood cell count of 13,800 with 17% polys , 0% bands . A repeat on day of life two , had a white count of 13,000 with 50% polys , 0% bands . A blood culture obtained prior to starting
antibiotics was no growth at 48 hours . 5. Gastrointestinal : Peak serum bilirubin occurred on day of life #4 , a total of 10.0/0.3 direct . A repeat bilirubin on 2012-05-19 was 8.0/0.3 direct . 6
Hematological : Hematocrit at birth was 43.8%. Raymond did not receive any transfusions of blood products during admission . 7. Neurology : Raymond has maintained normal neurological
examination throughout admission and there were no concerns at the time of discharge . 8. Sensory : Audiology hearing screening was performed with automated auditory brain stem
responses . Raymond passed in both ears .
</TEXT>
<TAGS></TAGS>
</ClinicalNarrativeTemporalAnnotation>

Figure 2: Input format schema

<EVENT id="E27" start="958" end="972" text="a burning pain" modality="FACTUAL" polarity="POS" type="PROBLEM"/>
<EVENT id="E28" start="995" end="1000" text="worse" modality="FACTUAL" polarity="POS" type="OCCURRENCE"/>
<TIMEX3 id="T7" start="1099" end="1104" text="10/92" type="DATE" val="1992-10" mod="NA"/>
<TIMEX3 id="T8" start="1259" end="1268" text="that time" type="DATE" val="1992-10" mod="NA"/>
<TIMEX3 id="T2" start="145" end="154" text="two years" type="DURATION" val="p2y" mod="NA"/>
<TIMEX3 id="T9" start="1577" end="1581" text="9/26" type="DATE" val="1993-09-26" mod="NA"/>
<TIMEX3 id="T10" start="1647" end="1656" text="this time" type="DATE" val="1993-09-29" mod="NA"/>
<TIMEX3 id="T0" start="18" end="28" text="09/29/1993" type="DATE" val="1993-09-29" mod="NA"/>
<TIMEX3 id="T11" start="2022" end="2044" text="the third hospital day" type="DATE" val="1993-10-01" mod="NA"/>
<TIMEX3 id="T12" start="2156" end="2163" text="10/2/93" type="DATE" val="1993-10-02" mod="NA"/>
<TIMEX3 id="T13" start="2249" end="2271" text="the day of discharge ." type="DATE" val="1993-10-04" mod="NA"/>
<TIMEX3 id="T14" start="2400" end="2405" text="10/92" type="DATE" val="1992-10" mod="NA"/>
<TIMEX3 id="T3" start="290" end="294" text="1991" type="DATE" val="1991" mod="NA"/>
<TIMEX3 id="T1" start="46" end="56" text="10/04/1993" type="DATE" val="1993-10-04" mod="NA"/>
<TIMEX3 id="T4" start="524" end="533" text="that time" type="DATE" val="1991" mod="NA"/>
<TIMEX3 id="T5" start="859" end="873" text="the past month" type="DURATION" val="p1m" mod="NA"/>
<TIMEX3 id="T6" start="920" end="929" text="last year" type="DURATION" val="p1y" mod="NA"/>
<TLINK id="TL0" fromID="E0" fromText="Admission" toID="T0" toText="09/29/1993" type="OVERLAP"/>
<TLINK id="TL1" fromID="E3" fromText="presented" toID="E0" toText="Admission" type="OVERLAP"/>

Figure 3: Input format for : EVENT, TIMEX3 and TLINK tags.

There exists an additional tag, namely SECTIME tag, as shown in figure 4, and matches two special temporal milestones in the summary of discharge, Admission or Discharge special events, so there usually exists an overlapping with TIMEX3 tags referred to those special event dates, usually identified by T0 and T1.

```
<SECTIME id="S0" start="18" end="28" text="09/29/1993" type="ADMISSION" dvalue="1993-09-29"/>
<SECTIME id="S1" start="46" end="56" text="10/04/1993" type="DISCHARGE" dvalue="1993-10-04"/>
```

Figure 4: Input format for the SECTIME tag.

In the following sections Events and Temporal expressions and Temporal relations we explain in depth the analysis of the prediction scenarios. The completion of those scenarios is required to solve the pipeline exposed in End to end task which main idea is output the provided (.xml) raw files with all the TIMEX3, EVENT, SECTIME and TLINK annotations.

2.1 Events and Temporal expressions

This is the first prediction scenario that consists in automatically tag those text chunks likely to be TIMEX3 or EVENT, then we have to predict each feature and ensemble all the information together. For modular purposes we differentiate the scenario into two isolated sub tasks: sequence labelling and attribute prediction. We explain in detail in the sections Sequence Labelling and Attribute prediction.

2.1.1 Sequence Labelling

In this case we aim to automatically annotate EVENT and TIMEX3 tags given certain sentence, this task is an extended version of the previous (Uzuner et al., 2011). In the following examples we indicate EVENT as "[text chunk]***EVENT***" and TIMEX3 as "[text chunk]***TIMEX3***":

The patient was [transfused]***EVENT*** with [radiated Leuco]***EVENT*** three units .

Dr. Lawrence performed [the surgery]***EVENT*** .

His hematocrit was checked [two weeks]***TIMEX3*** prior to admission .

She reports a 5 pound weight loss over [the past several months]***TIMEX3***.

The patient was brought to the operating room on [05-04-1998]***TIMEX3*** .

Applying sequence labelling techniques we propose to annotate the chunks, then those chunks are used for attribute prediction purposed as described in the section Attribute prediction. In addition we have to delimit the text in between the proper relative start and end offsets. All this facts encourages the requirement of clear and non-ambiguous re construction way of the data, we also require a modular design to deal with the data due labelled sequences and the offsets are used for assemble final tags.

2.1.2 Attribute prediction

EVENT tag marks the events described in the medical record that are significant to the patient's clinical timeline, we indicate them in the following examples as "[event text chunk]***Tags***". EVENTS have one of the following types: PROBLEM, TREATMENT, TEST, CLINICAL_DEPT, OCURRENCE or EVIDENTIAL. Below are some examples:

The patient was [transfused]***TREATMENT*** with [radiated Leuco]***TREATMENT*** three units .
The patient was previously [admitted]***OCURRENCE*** for [cardiac and pulmonary disease]***PROBLEM*** .
Dr. Lawrence performed [the surgery]***TREATMENT*** .

In addition there exist other attributes as polarity: POS or NEG. And modality: PROPOSED, FACTUAL, CONDITIONAL or POSSIBLE. The annotated tags assemble type, modality and polarity as shown in the following example:

This is very likely to be [an asthma exacerbation]***PROBLEM POSSIBLE NEG***.

In the following figures 5, 7 and 6 we can see the distribution for type, polarity and modality distributions over training set. As we can see, there exists high imbalance in the three attributes, this leads us to develop a single expert system for each of the features.

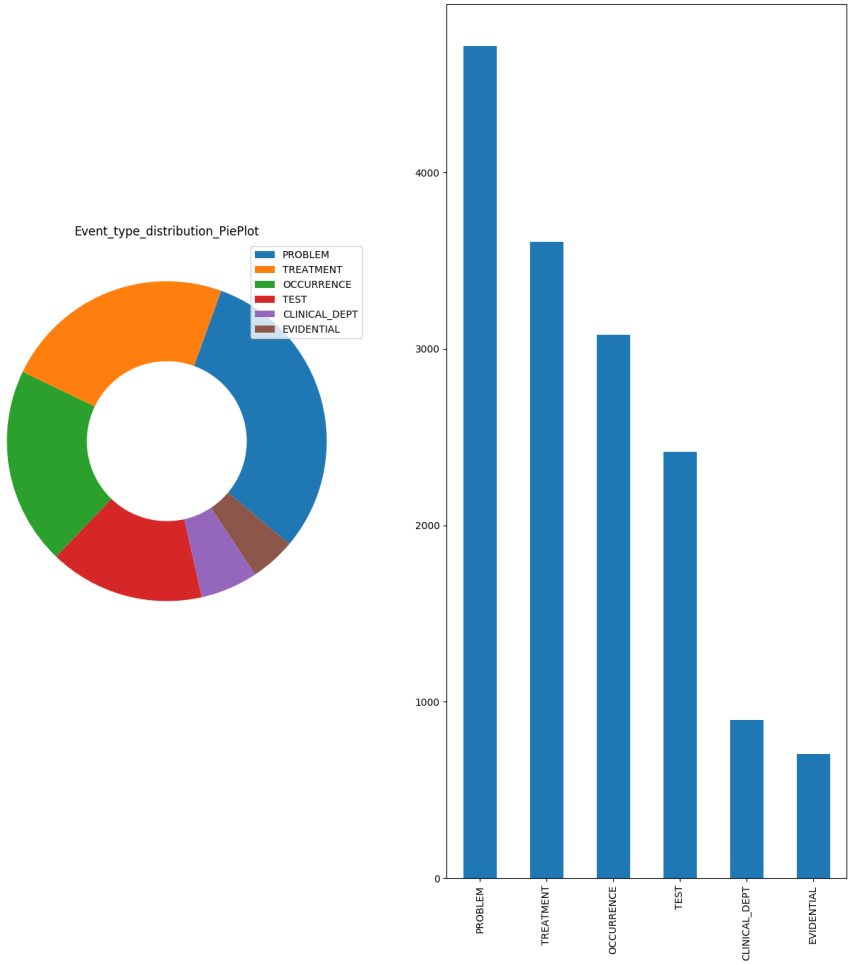

Figure 5: Distribution for EVENT-type in train set.

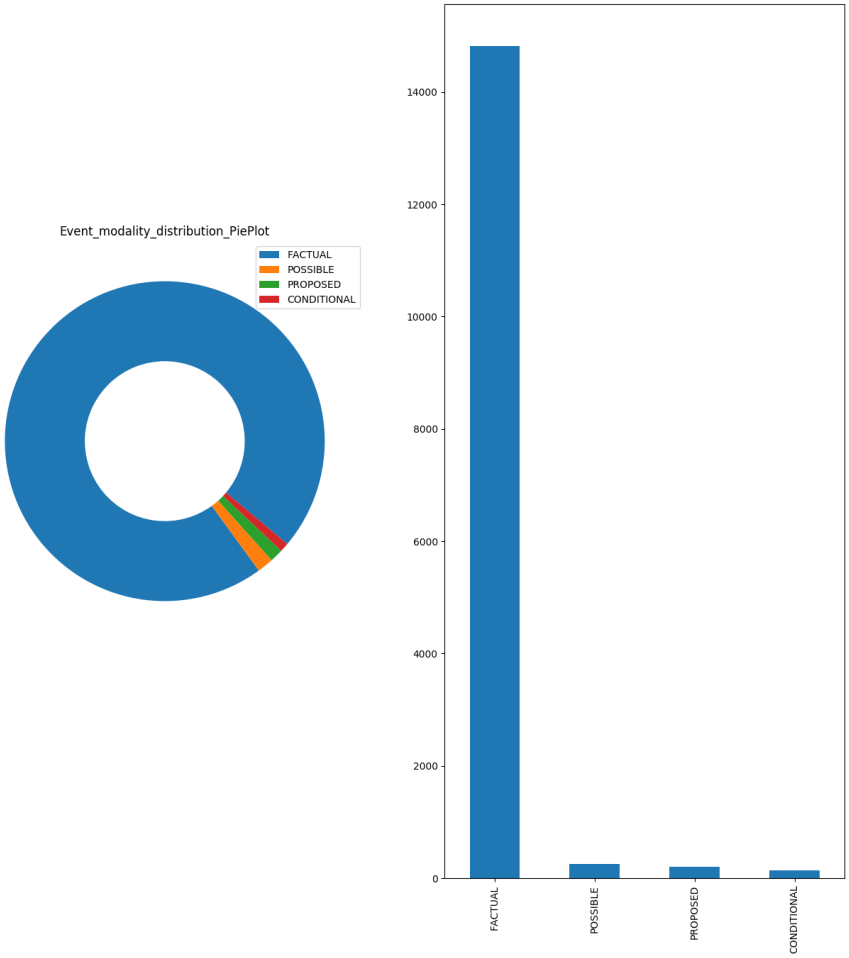

Figure 6: Distribution for EVENT-modality in train set.

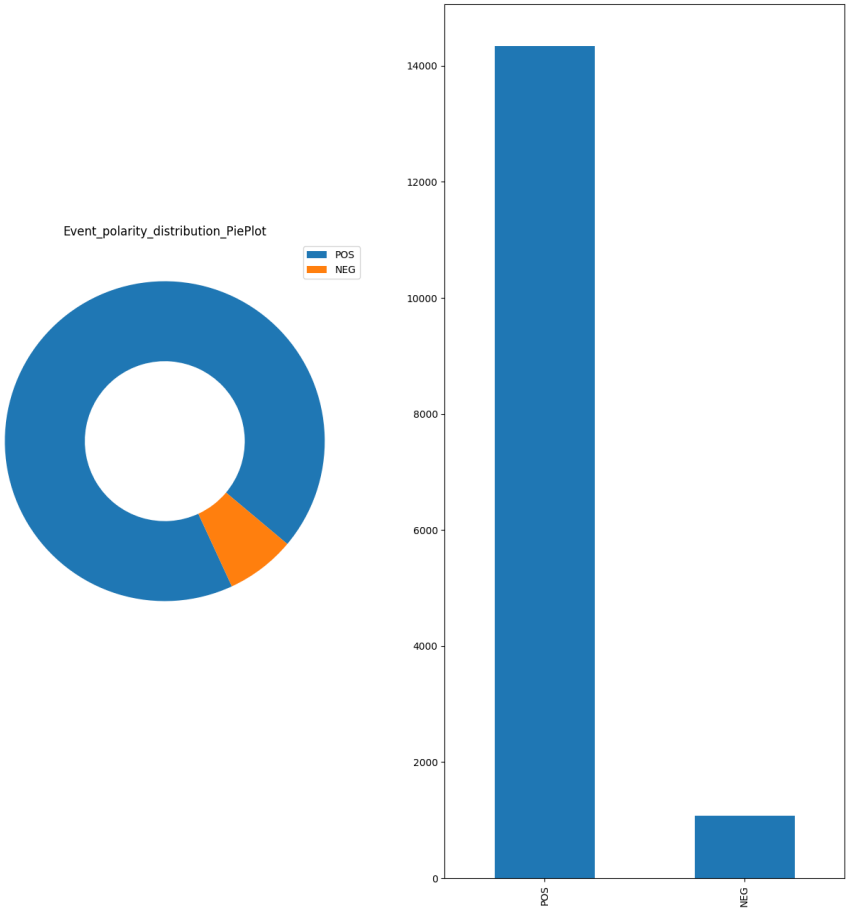

Figure 7: Distribution for EVENT-polarity in train set.

TIMEX3 tag marks temporal expressions that include all references to: points in time, time periods, durations, and frequencies, we indicate them in the following examples as "[timex3 text chunk]***Tags***". TIMEX3 have one of the following types: DATE, TIME, DURATION or FREQUENCY. Below are some examples:

His hematocrit was checked [two weeks]***FREQUENCY*** prior to admission .
She reports a 5 pound weight loss over [the past several months]***DURATION***.
The patient was brought to the operating room on [05-04-1998]***DATE*** .

In addition there exist other attributes as modifier: NA, APPROX, END, START, MORE or MIDDLE. And value, this isn't nominal and consists on the ISO:8601 transcription of the tag text. TIMEX3 follow THYME project (Styler IV et al., 2014) guidelines. The annotated TIMEX3 tags assemble type, modifier and value attributes as shown in the following example:

The patient was brought to the operating room on [05-04-1998].***DATE NA Value="1998-05-04"***.

In the following figures 8 and 9 we can see the distribution for type and modifier over training set. There exists high imbalance in the two attributes, this leads us to develop a single expert system for each of the attributes as we have done with EVENT tag.

In the case "value" attribute, the task involves research on parsing and normalization into ISO:8601 standard, in this case the difficulty besides to the ambiguity of time expressions, some require context information, and specific modules to deal with Frequency and Duration, for example:

"02-21" $\longrightarrow$ "1991-02-21"

"few weeks" $\longrightarrow$ "P3W"

"next Monday" $\longrightarrow$ "2019-06-22"

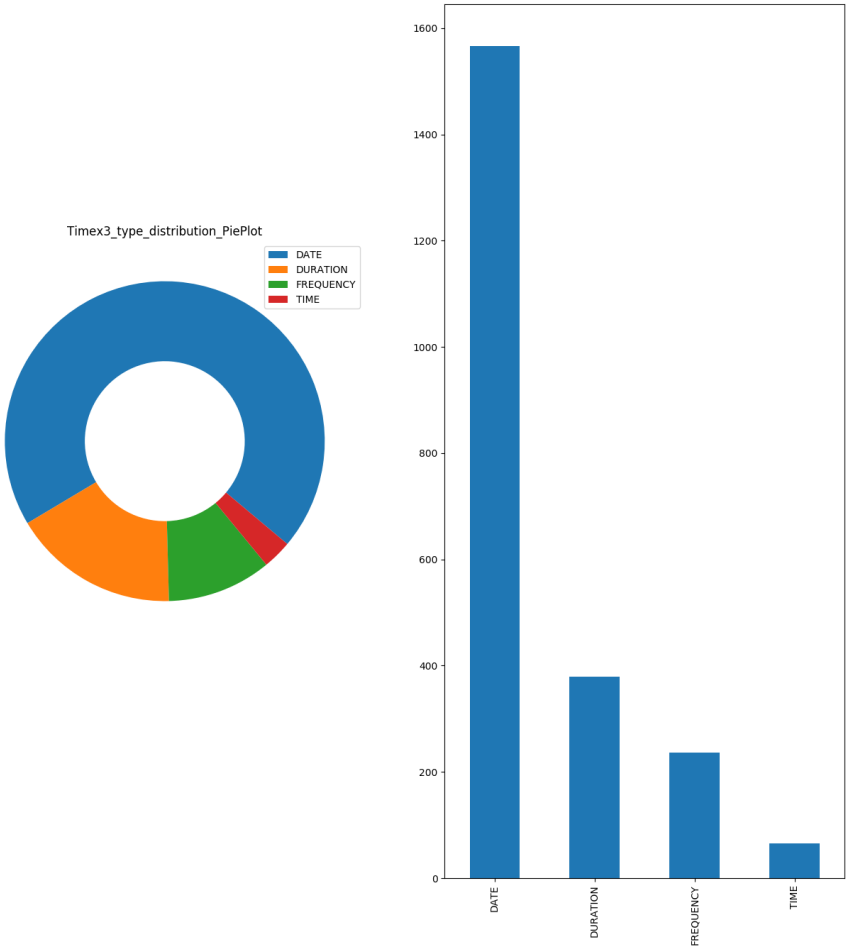

Figure 8: Distribution for TIMEX3-type in train set.

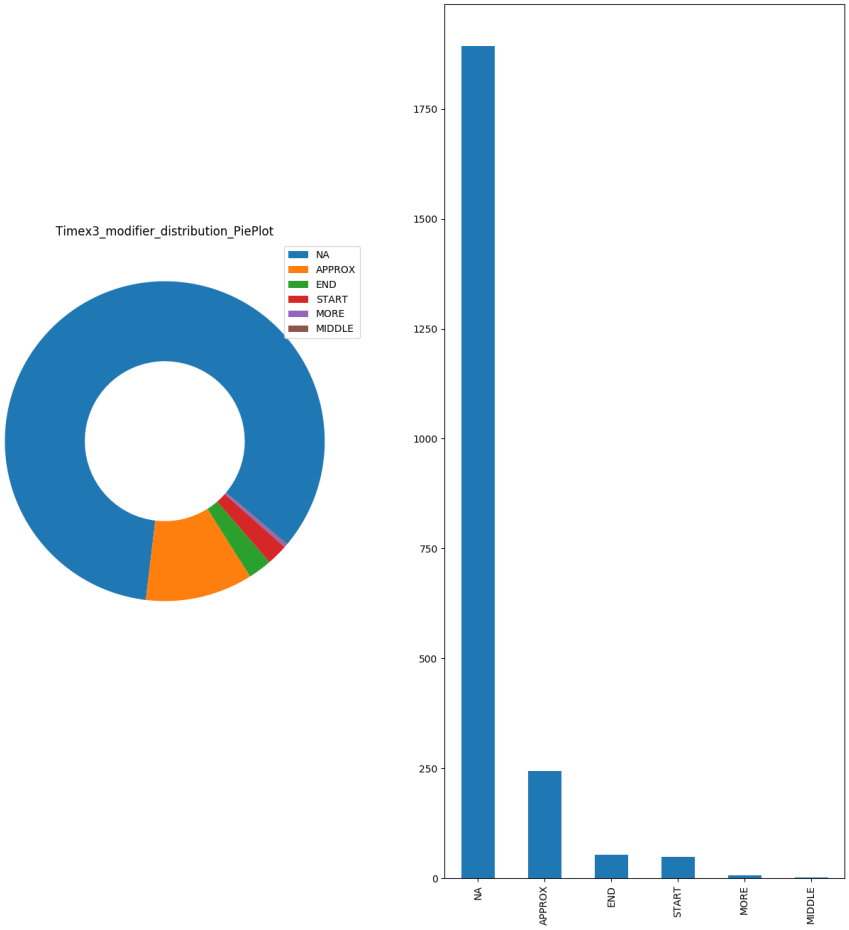

Figure 9: Distribution for TIMEX3-modifier in train set.

2.2 Temporal relations

This is the second prediction scenario that consists in to automatically detect relations between the annotated tags applying supervised classification techniques, in this case the classification determines whether certain relation exists between the entities, and which is the relation. TLINK task aims to develop a suitable relation extraction system following the constraints of the task exposed in the Relation Extraction Scenario description.

2.2.1 Relation Extraction Scenario description

In this task we aim to automatically detect relations among tags, the relation is defined by its type: OVERLAP, BEFORE or AFTER. In this way we detect the temporal relation distribution, relations are expressed by TLINK tags, those are composed by fromID, toID, fromText and toText attributes. So, the task comprises the detection of the relation and the composition of the TLINK Tag.

In this task we have to initially define the way we relate 'from' tags with 'to' tags in order to check relations, the direction is important because the direction of the relation is required to describe the chronological order of the tags in the clinical history, and the additional requirement to describe the chronological order of the tags referred to the general historical information of the patient, this last one is handled by the SECTIME data, the 'admission' and 'discharge' reference. We also have to deal with long-dependency tracking referred to inter-sentence relations and short dependency tracking referred to intra-sentence relations. This lead us to require the design of synthetic data set to learn 'relation' vs 'no relation' categories. In other words, we decide how to pairwise the tags, in this way we define the detection of the relation.

Here we provide some insight in the following exaples about the targets: EVENT-SECTIME, EVENT-EVENT, EVENT-TIMEX3, TIMEX3-EVENT and TIMEX3-TIMEX3 respectively. Each target follow different direction patterns and context length, we denote [from entity]***e1***, [to entity]***e2***, context as (...) and relation ***TYPE*** :

[Admisssion]***e1*** ... [13-02-1997]***e2*** ... ***OVERLAP***

[Admisssion]***e1*** ... [showed hypertension]***e2*** ... ***OVERLAP***

[moved to surgery]***e1*** ... [third admittance day]***e2*** ... ***OVERLAP***

[At that time]***e1*** ... [was scheduled]]***e2*** ... ***AFTER***

[13-02-1997]***e1*** ... [At that time]***e2*** ... ***OVERLAP***

The challenge on this task consist on facing the high imbalance of the synthetic data, because in the definition we have to take into account the target relation, to define the suitable direction and range of dependency to track. Handling the wider range of relations (all for example) requires to generate a huge number of empty relations (in between all tags following the example). We must fit the process into reasonable computational costs.

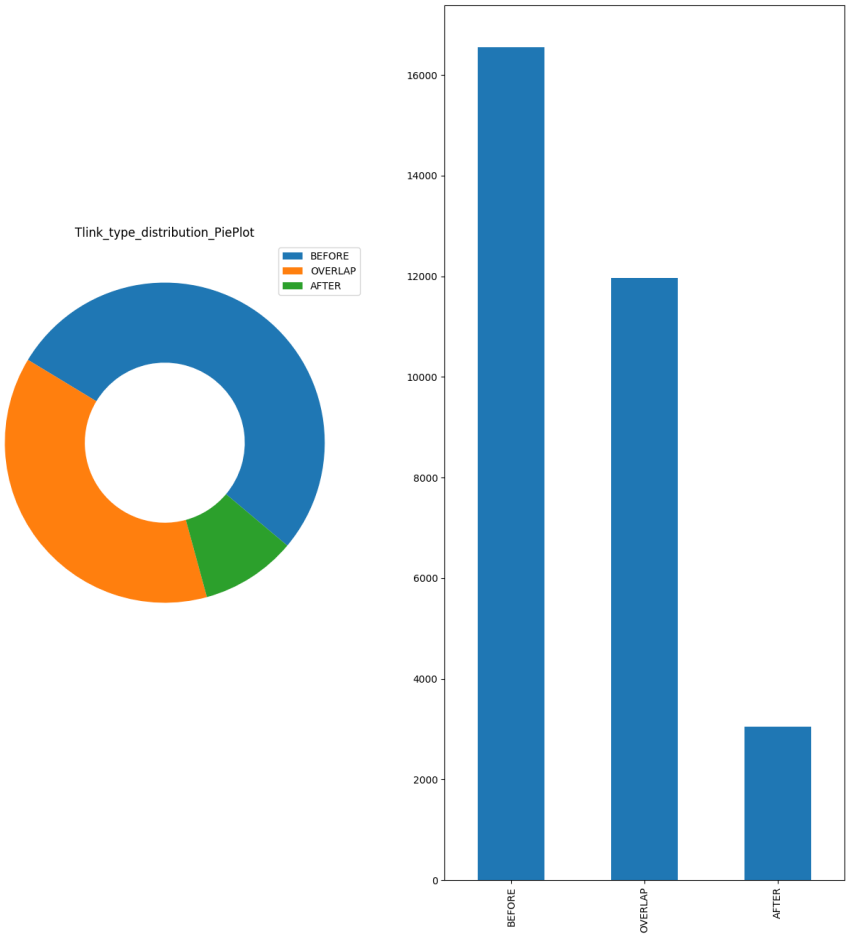

Figure 10: Distribution for TLINK-type in train set.

2.2.2 End to end task

This is the final scenario that consists in to automatically detect the relations among the EVENT and TIMEX3 tags starting off raw clinical texts. The scenario conjugates Events and Temporal expressions and Temporal relations into what we understand as a single pipeline to track the overall performance of the system.

In this task we have to ensure good performance in the Sequence Labelling task to minimize the error propagation into the relation extraction task, in addition we have to deal with Attribute prediction in order to reach final (.xml) files fully annotated. Intermediate formats we use across the process are relevant in order to achieve a whole modular system that allow us perform isolated tasks and even the entire pipeline indistinctly.

We have to fit the pipeline computational into affordable costs, taking into account that system must process certain amount of files at a time. In the section System Description we define the suitable designs for the problem, in the next section State of the Art we introduce the information taken into account for Design purposes.

3 State of the Art

In the section Overview we explain the state of the art technology for domain specific sequence labelling and relation extraction, and we provide an insight of previous end-to-end approaches to solve the i2b2 challenge. Then in the section Evaluation we explain the evaluation metrics applied for "i2b2 2012 challenge" (Sun et al., 2013).

3.1 Overview

Medical sequence labelling, as opposed to sequence labelling, shows certain specificities (Zhou et al., 2004), like their descriptive nature, their productivity and the massive use of acronyms. These specificities and the fact that static embeddings were systematically employed by Sequence Labelling systems, yield researchers to use in-domain corpus, as opposed to general-domain corpus, to both train the Medical Sequence Labelling systems as well as the static pre-trained embeddings, in order to obtain better results since controlling domain leads to better control on polysemy ((Soares et al., 2019), (Stenetorp et al., 2012)). Recently, performance of both Sequence Labelling and Medical Sequence Labelling tasks has shown a significant breakthrough with the introduction of contextualized word embeddings (ELMo (Peters et al., 2018), ULMFiT (Howard and Ruder, 2018), BERT (Devlin et al., 2018a) and FLAIR (Akbik et al., 2018a)). Although contextualized embeddings seem to reduce the gap between general and domain specific corpus, several works on Medical Sequence Labelling task argue that domain-specific contextualized embeddings still yields superior performance over the standard and general-domain word embeddings ((Akhtyamova et al., 2020),(Lee et al., 2019),(Si et al., 2019),(Sheikhshabbafghi et al., 2018)).The present Medical Sequence Labelling task due to its heterogeneity (`EVENT` are more specific to the medical domain while `TIMEX3` are less specific) represents a perfect task to check the performance of contextualized Language Models.

Recently, transfer learning has shown to be a successful alternative when no annotated data is available in the target domain and language (Devlin et al., 2018a; Baldini Soares et al., 2019). Recent Transformer sequence models (Vaswani et al., 2017) outperformed the state-of-the-art in relation extraction (Baldini Soares et al., 2019; Peters et al., 2019; Joshi et al., 2019). Some works try to integrate knowledge bases into Transformer sequence models (Peters et al., 2019). Nevertheless, simpler approaches based on entity markers, that consist on mark in the input sentence the entities involved, show same competitive performance with a quicker setup. In a similar manner, multilingual language models (Lample and Conneau, 2019) have shown impressive capacity to perform zero-shot learning in Cross-lingual tasks. This kind of models seems very promising for relation extraction tasks with small annotated training set as we have in i2b2 challenge.

We explored the FLAIR framework (Akbik et al., 2018b) in our eHealthKD 2020challenge approach (Andrés et al., 2020), the eHealthKD 2020 challenge (Piad-Morffis et al.) consists on a similar Sequence labelling and relation extraction applied over Spanish clinical texts, in this case we submitted a Sequence Labelling approach based on FLAIR system showing high improvements in relation to the Lample tagger (Lample et al., 2016) framework, the architecture is very similar but in FLAIR system we have support for contextual embeddings, in this case we can decide the WE composition to represent the context data adequately via static WE and dynamic character based WE as we can see in the figure 11, this feeds the output BiLSTM + CRF architecture. The appearance of BERT (Devlin et al., 2018b) changed the way we understand LM techniques, in this case a separated deep neural network represents the context data and then decisions are inferred from the fine-tuning over detailed textual representation as shown in 12.

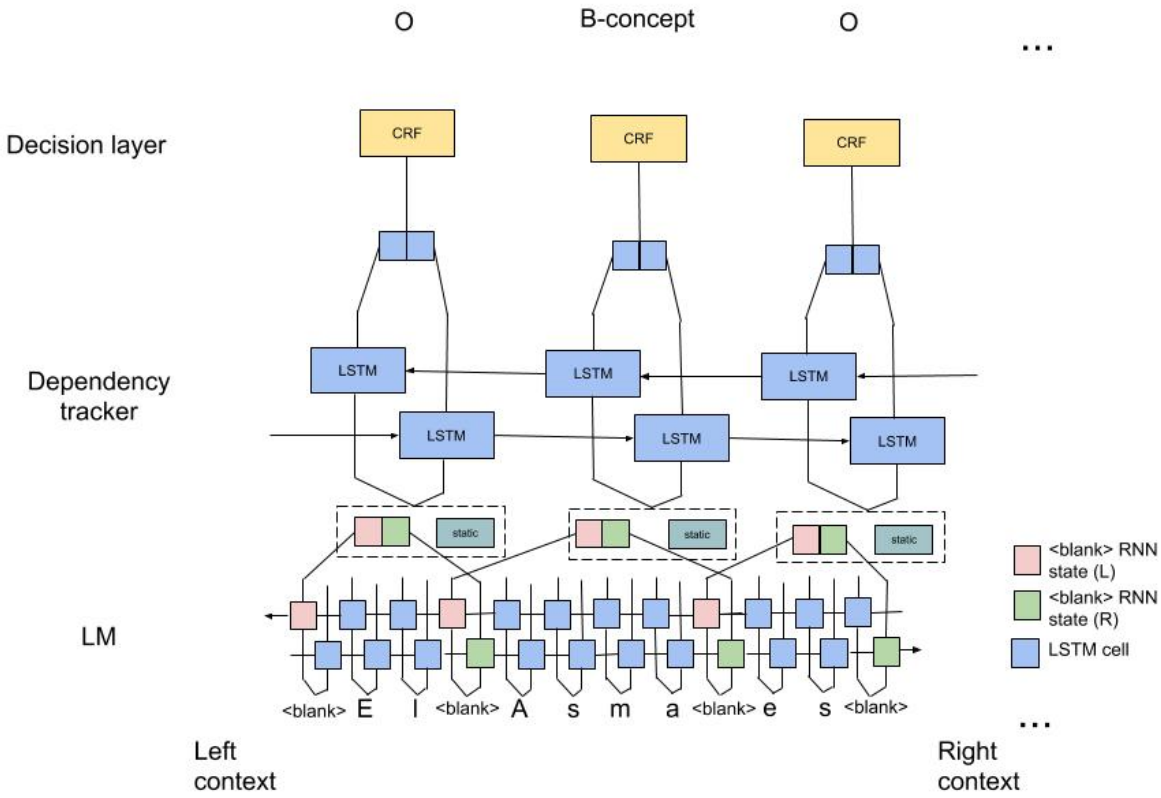

Figure 11: Architecture of FLAIR system.

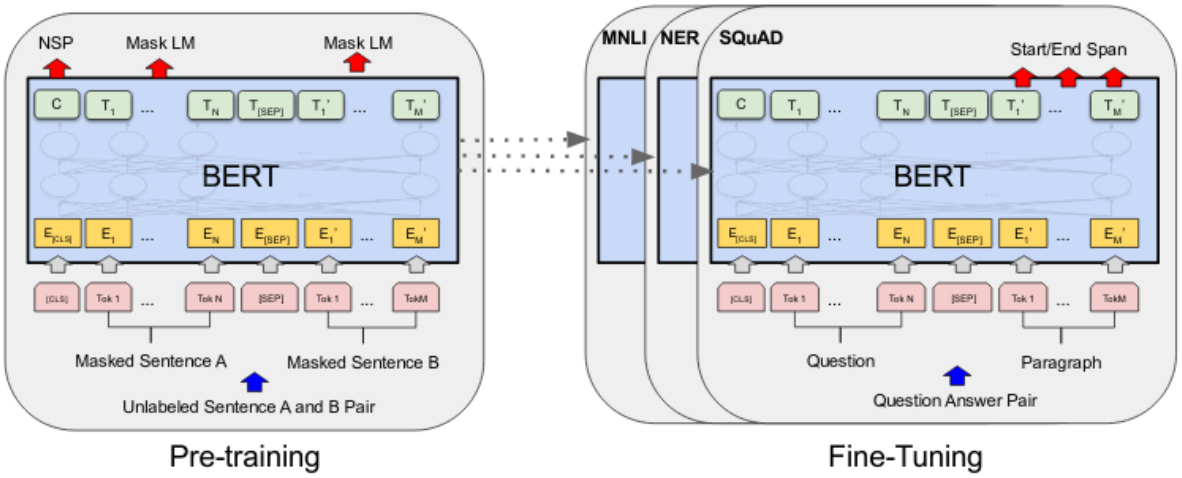

Figure 12: Architecture of BERT system.

BERT architecture improvements lead to Cross-lingual models (Conneau et al., 2019a) as XLM-RoBERTa (Conneau et al., 2019b) appeared, those are pre-trained with huge amounts of data in wide range of languages ,provided by Transformers framework (Wolf et al., 2019) improved dramatically the LM properties being able to transfer learning between languages only requiring domain adaptation in the target language, in addition the Transformers framework allows to use the most powerful Deep Learning frameworks such as Tensorflow (Abadi et al., 2015) and PyTorch (Paszke et al., 2019) with the SOA LM solutions. In this way we are allowed to use linear layers as CRF or non-linear layers as Softmax in the model we use for Sequence Labelling and Supervised Classification tasks. We compared FLAIR in front of XLM-RoBERTa base, and XLM-RoBERTa base outperformed FLAIR by 3 F1 score points.

Finally we briefly introduce the main features of the best state of the art end-to-end systems for i2b2 2012 challenge (Sun et al., 2013)[Table 4], those proposed by Vanderbilt University and Beihang University;Microsoft Research Asia,Beijing; Tsinghua University respectively. The first approach consists on: sequence labelling using CRF and attribute prediction using SVM for EVENT extraction, Rule based sequence labelling and attribute prediction using HeidelTime for TIMEX3 extraction and Rule based pair selection+CRF data generation with relation extraction using SVM for TLINK extraction. The second approach consists on: sequence labelling using and attribute prediction using CRF for EVENT extraction, CRF based sequence labelling and attribute prediction using SVM + Rule based for TIMEX3 extraction and relation extraction using SVM for TLINK extraction.

3.2 Evaluation

In the challenge we are provided evaluation scripts to guarantee the comparison between submissions. In (Sun et al., 2013)[Methods] we can see the way this software evaluates submissions. Those metrics are used in order to compare our system with previous approaches in Experiments section.

EVENT extraction task span detection and identifying their attributes. They used the F measure, the harmonic mean of precision and recall of the system output against the gold standard to evaluate EVENT span detection performance.

TIMEX3 extraction requires span detection, attribute identification, and value normalization. TIMEX3s values attributes need to be normalized to ISO:8601 standards. They used the F measure of the TIMEX3 span detection multiplied by the accuracy of the value field as the primary metric for TIMEX3 extraction evaluation.

TLINK extraction uses F measure as the primary evaluation metric. Prior to evaluation, they compute the temporal closure of the TLINKS provided by the system and the temporal closure of the TLINKs found in the gold standard. The precision of the system output is the percentage of system TLINKs that can be verified in the temporal closure of the gold standard TLINKs. Recall of the system output is the percentage of gold standard TLINKs that can be verified in the temporal closure of the system TLINK output. They calculate the temporal closure based on TempEval3 method (UzZaman and Allen, 2011).

4 System Description

As stated previously, XLM-RoBERTa has shown to obtain good results in several tasks. XLM-RoBERTa is a combination of RoBERTa and the updated version of the original XLM model. Both the sentences and words of the training data set are partitioned into most used subwords in the corpus of all training languages. XLM-RoBERTa improves XLM since it uses a larger corpus (2.5TB) and 100 languages for training instead of 15. The second component, RoBERTa, is a robust version of BERT but as with the first component, this one too has been trained on a larger corpus (Joshi et al., 2019). Unlike mBERT who has been trained on Wikipedia, XLM-RoBERT uses the CommonCrawl(Conneau et al., 2019a) corpus for its training. In this section we explain step by step our approach. First we adapt the XLM-RoBERTa system to the clinical domain in Language Model adaptation section, then in EVENT and TIMEX3 task section we explain sequence labelling and attribute prediction processes, in TLINK task we explain the relation extraction process, finally in **??** section we explain the pipeline assembling.

4.1 Language Model adaptation

Cross-lingual models seem to be language-dependant according to (Podolak and Zeinert, 2020) to adapt the cross lingual language model to specific domains we apply the process shown in figure 13.

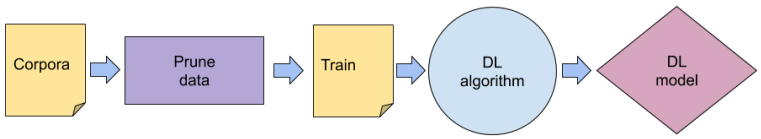

Figure 13: Language Model Adaptation schema.

Pre-process: We generate the training data set with the desired amount of sentences from a given corpora, the data set is handled in a (.txt) file with a sentence per row.

Training: For domain adaptation we fine tune the generic XLM-RoBERTa tokenizer with the generated data, this process fine tune the tokenizer to achieve the lowest perplexity according to the proposed data.

4.2 EVENT and TIMEX3 task

Both systems shown in figures 14 and 15 aim to automatically extract EVENT and TIMEX3 tags. Textual expressions are extracted using sequence labellers based on previously adapted XLM-RoBERTa-base tokenizer, each of the labellers are fine tuned, one in EVENT extraction and the other in TIMEX3 extraction, the main idea is to retrieve the textual spans at this stage. Each textual expression is classified without regarding context in independent sentence classifiers based on previously adapted XLM-RoBERTa-base tokenizer, one fine tuned classifier for each attribute prediction sub task. Finally the results of sequence labellers and attribute predictions are assembled into the final tags.

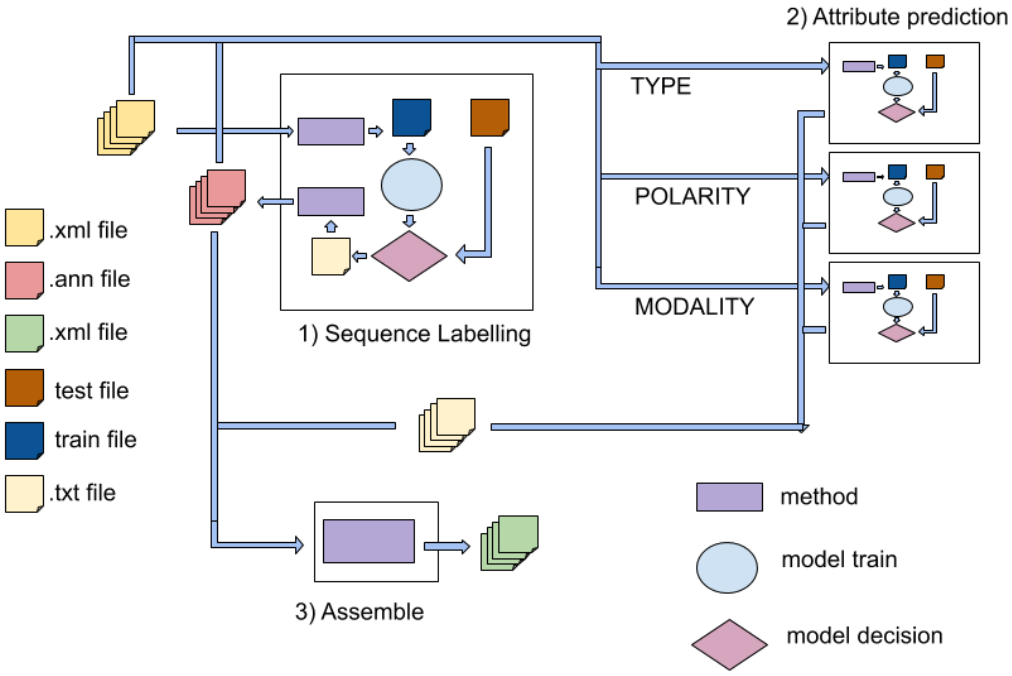

Figure 14: EVENT system schema.

In the particular TIMEX3 system shown in 15 we introduce an ISO:8601 parsing module based on regular expression, this module aims to normalize time expressions.

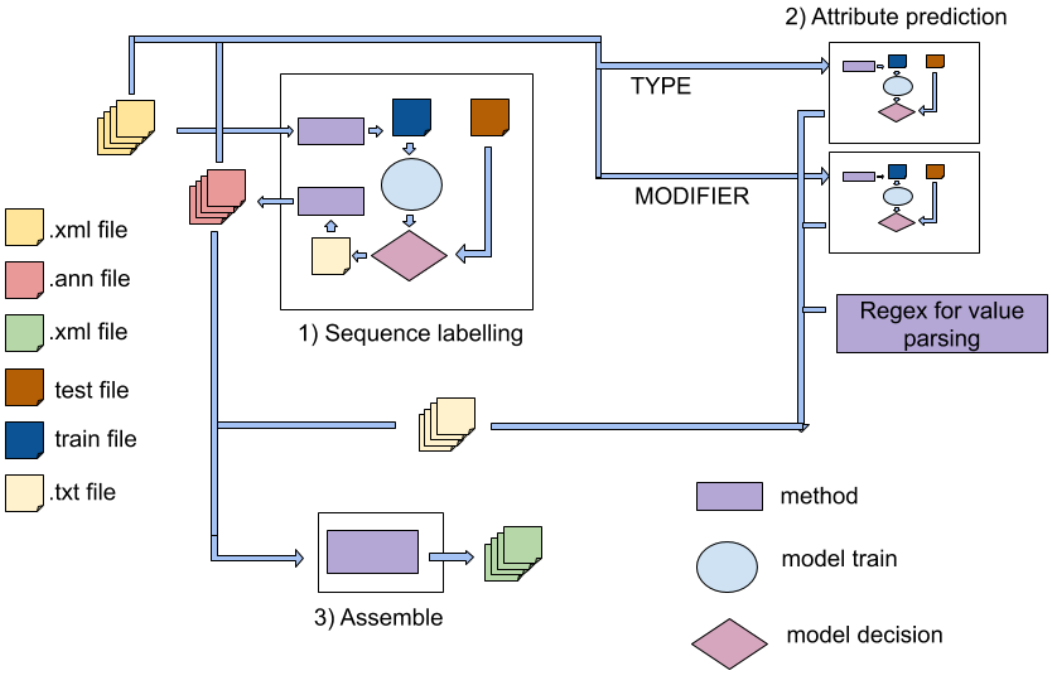

Figure 15: TIMEX3 system schema.

4.2.1 Sequence Labelling

To perform this task we process the data into (.ann) format able to track textual spans, we apply the process shown in figure 16 for the desired Sequence Labelling task, TIMEX3 or EVENT.

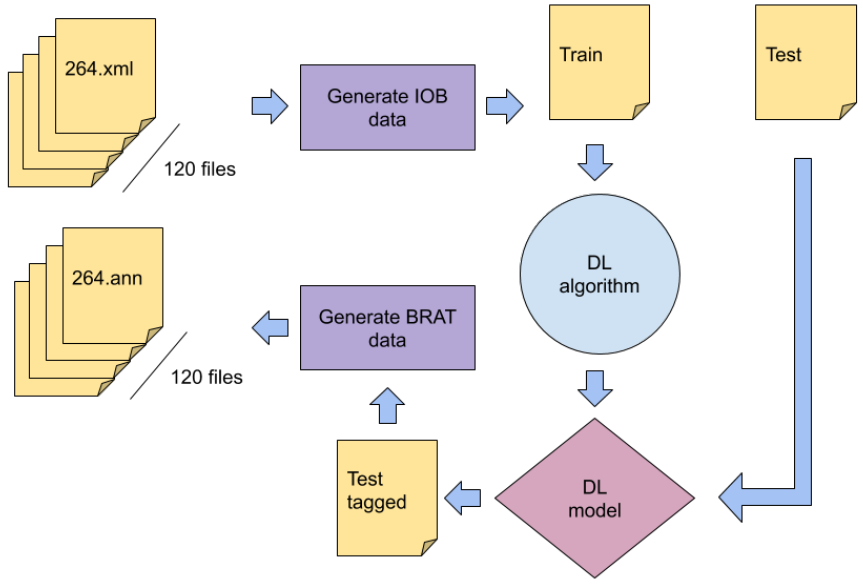

Figure 16: Sequence Labelling process schema.

Pre-process: We generate meta data schema as shown in the following figure 17 to allow locating each label unequivocally using a defined pattern start ('$') and end ('€'), this allow disambiguation for repeated entities.

```
$ EVENT ADMISSION € DATE :
10/19/93
$ EVENT DISCHARGE € DATE :
10/25/93
HISTORY OF PRESENT ILLNESS :
Mrs. Dua is a 34 year old white female with an unclear history of $ EVENT hypertension € for at least five years .
```

Figure 17: meta data format.

Tokenization is performed splitting the blanks in meta data format and sentences divided are divided splitting the line breaks in meta data, this information is used to generate the IOB format as shown in figure 18 placing a token per row, introducing a blank between sentences and two blanks between documents. The IOB format is used to fine tune the sequence labellers.

```
ADMISSION B-EVENT
DATE O
: O

10/19/93 O

DISCHARGE B-EVENT
DATE O
: O

10/25/93 O

HISTORY O
OF O
PRESENT O
ILLNESS O
: O

Mrs. O
Dua O
is O
a O
34 O
year O
old O
white O
female O
with O
an O
unclear O
history O
of O
hypertension B-EVENT
for O
at O
least O
five O
years O
. O
```

Figure 18: IOB format.

Training:For Sequence labelling purpose we used 'XLMRobertaForSequenceClassification' (Alexis Conneau and Stoyanov, 2020) models, the IOB format is feed to the domain adapted 'XLMRobertaTokenizer' (Alexis Conneau and Stoyanov, 2020), the system generates subword representation between [CLS] and [SEP] special tokens, internally maps all the information into the cross lingual language model vector space, for each input token the corresponding label is classified via Softmax linear layer is done in 'TokenClassifierOutput' (Alexis Conneau and Stoyanov, 2020).

Post-process: As we require text spans as result, we convert IOB format into (.ann) format as shown in figure 19.

```
T1      EVENT 1 10      Admission
T2      EVENT 29 38     Discharge
T3      EVENT 156 168   hypertension
```

Figure 19: (.ann) format.

4.2.2 Attribute prediction

To perform this task we process the data into tabbed format able to track the textual expression category, we apply the process shown in figure 20 for the desired Attribute prediction task.

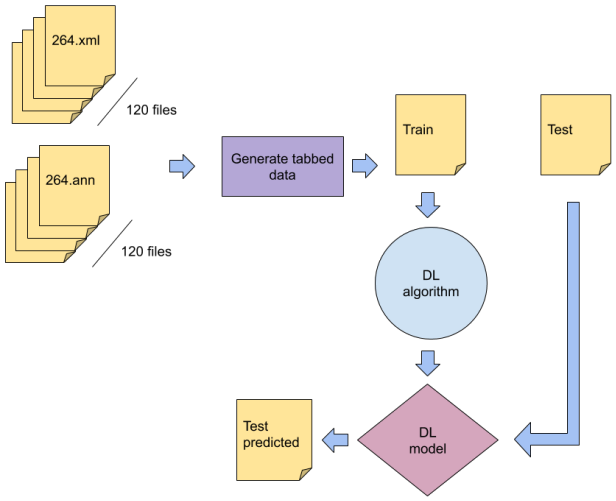

Figure 20: Attribute process schema.

Pre-process: We generate a tabbed format with two columns: the textual expression and the category as shown in the figure 21.

```
Admission       OCURRENCE
Discharge       OCURRENCE
hypertension    PROBLEM
```

Figure 21: tabbed format.

Training:For Attribute classification purpose we used 'XLMRobertaForSequenceClassification' (Alexis Conneau and Stoyanov, 2020) models, the tabbed format is feed to the domain adapted 'XLMRobertaTokenizer' (Alexis Conneau and Stoyanov, 2020), the system generates subword representation between [CLS] and [SEP] special tokens, internally maps all the information into the cross lingual language model vector space, [CLS] token is used to classify each textual expression label via Softmax linear layer done in 'SequenceClassifierOutput' (Alexis Conneau and Stoyanov, 2020).

4.2.3 Assemble

Finally we have to assemble textual spans with each predicted attribute label, we apply the process shown in figure 22.

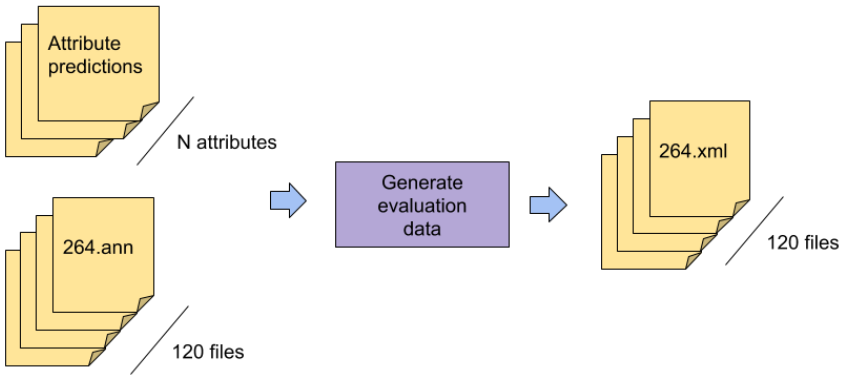

Figure 22: Assemble process schema.

Post-process: We generate the final tags as shown in figure 23, we extract: the offsets, textual expression, id and tag; from (.ann) format, and we extract the attribute categories from tabbed format.

<EVENT id="E27" start="958" end="972" text="a burning pain" modality="FACTUAL" polarity="POS" type="PROBLEM"/>

.ann tabbed format

Figure 23: Assembled tag example.

4.3 TLINK task

The system shown in figure 24 aim to automatically extract relations among EVENT and TIMEX3 tags. EVENT and TIMEX3 tags are provided applying EVENT and TIMEX3 task or via EVENT / TIMEX3 gold standard depending on the selected task. Each textual expression is classified regarding context in independent sentence classifiers based on previously adapted XLM-RoBERTa-base tokenizer, one fine tuned classifier for each target prediction sub task. In this case we only designed the wider targets: Event-Sectime, Event-Event and Event-Timex3.

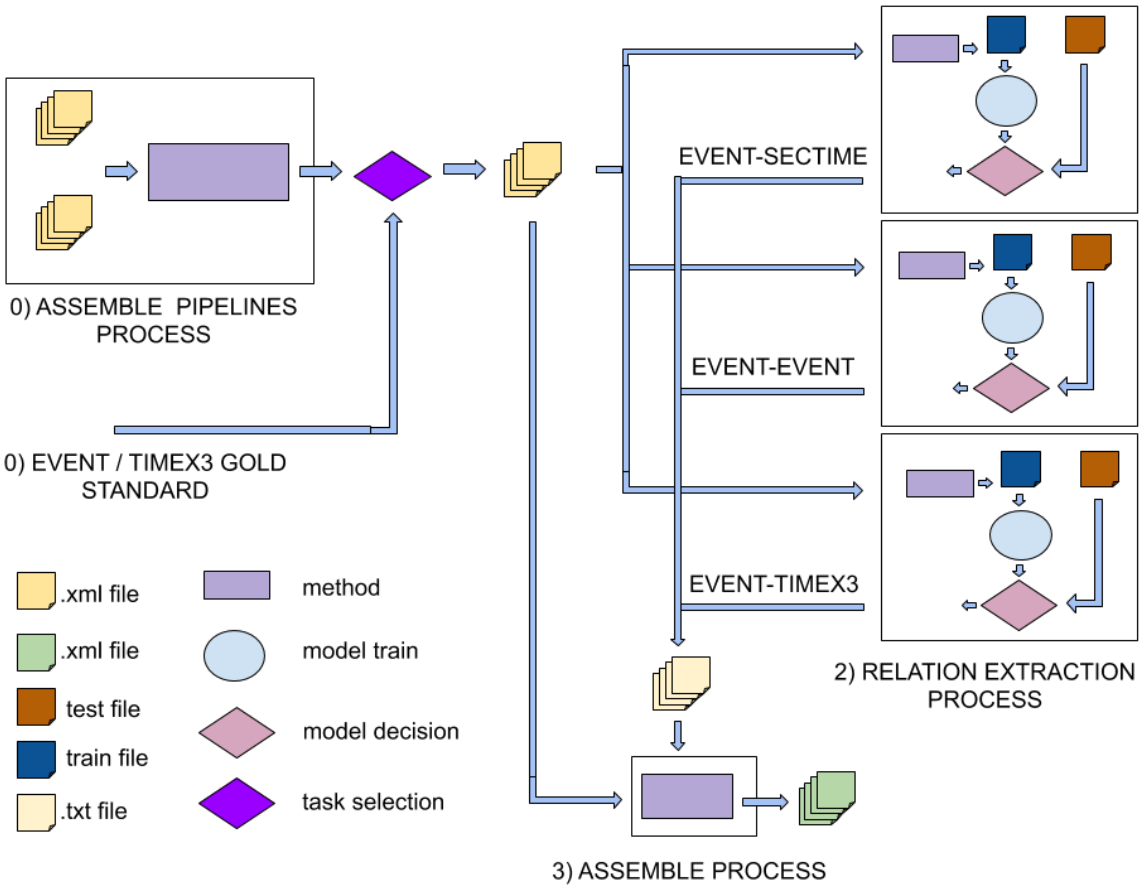

Figure 24: TLINK system schema.

4.3.1 Sub task selection

For TLINK extraction we have to decide whether we want to test the extraction performance over the EVENT / TIMEX3 gold standard or end-to-end pipeline, in this last case we have to assemble the EVENT and TIMEX3 results from the EVENT and TIMEX3 task systems as shown in the next figure 25, in this case we merge the tags of both systems into new (.xml) files similar to Event/Timex3 Gold standard.

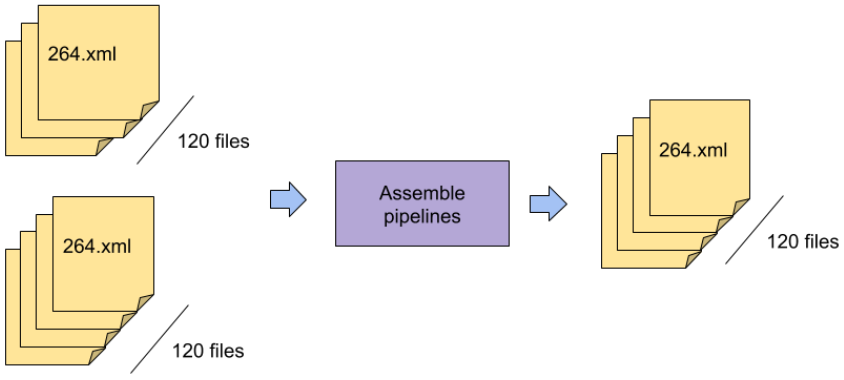

Figure 25: Assemble pipelines process schema.

4.3.2 Target Prediction

To perform this task we process the data into tabbed format able to track the textual expression category with context, we apply the process shown in figure 26 for the desired target prediction task.

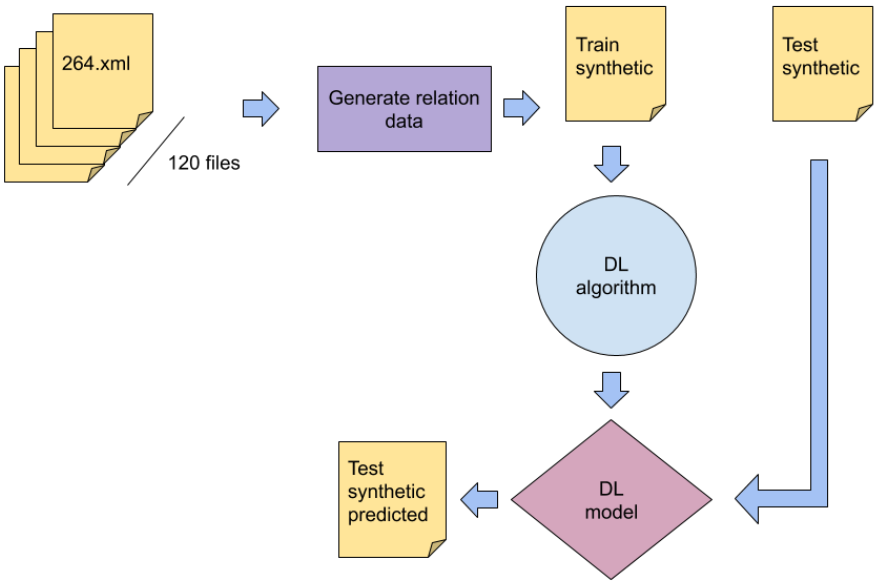

Figure 26: Target prediction schema.

Pre-process: We generate a tabbed format with two columns: the textual expression with context, entities are marked (Baldini Soares et al., 2019) in the context using [E1S], [E1E], [E2S] and [E2E] tokens for entity1 and entity2 respectively,and the relation type as shown in the figure 27.

```
[E1S] headache [E1E] , [E2S] diplopia [E2E]     OVERLAP
[E2S] A Memorial Hospital [E2E] where [E1S] her blood pressure [E1E]    OVERLAP
[E2S] Ro Woodma Healthcare [E2E] for [E1S] further management [E1E]     OVERLAP
```

Figure 27: Training data for TLINK example.

Training: For Target classification purpose we used 'XLMRobertaForSequenceClassification' (Alexis Conneau and Stoyanov, 2020) models, the tabbed format context with entity markers is feed to the domain adapted 'XLMRobertaTokenizer' (Alexis Conneau and Stoyanov, 2020) (markers are also added to tokenizer vocabulary), the system generates subword representation between [CLS] and [SEP] special tokens, internally maps all the information into the cross lingual language model vector space, [CLS] token is used to classify each target label via Softmax linear layer done in 'SequenceClassifierOutput' (Alexis Conneau and Stoyanov, 2020).

4.3.3 Assemble

Finally we have to assemble from / to textual spans and its associated identifiers with each predicted attribute label, we apply the process shown in figure 28.

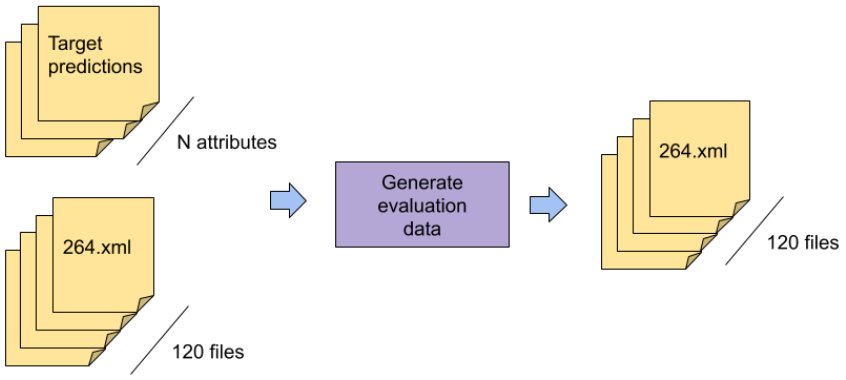

Figure 28: Assemble TLINK schema.

Post-process: We generate the final tags as shown in figure 29, we extract: the fromText, toText, id, fromId, toId and tag; from (.xml) format, and we extract the relation type from target predictions.

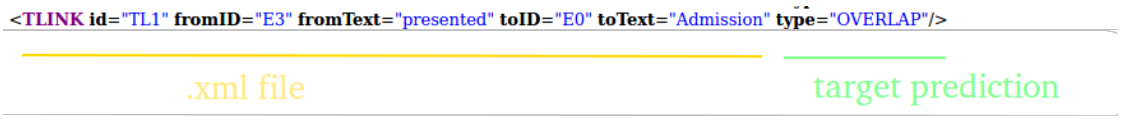

Figure 29: Assemble example for TLINK tag.

5 Experiments

In this section we explain the experiments performed in the current work, the experimental settings used and the achieved results.

5.1 Previous Approach

The EVENT and TIMEX3 task was performed in a previous approach (Andrés Santamaría, 2019) based on Bi-directional Long Short Tern Memory (BiLSTM) with Conditional Random Field (CRF) layer on top (Lample et al., 2016) as shown in the figure 30. This solution perform Accuracy (87.65 %) , precision (72%) , recall (76.68%) and F1 (74.27%) for EVENT-type extraction . And Accuracy (98.55%) , precision (81.58%) , recall (78.48%) and F1 (80%) for TIMEX3-type extraction. Even the results are not comparable, is a good starting point to understand current experiments.

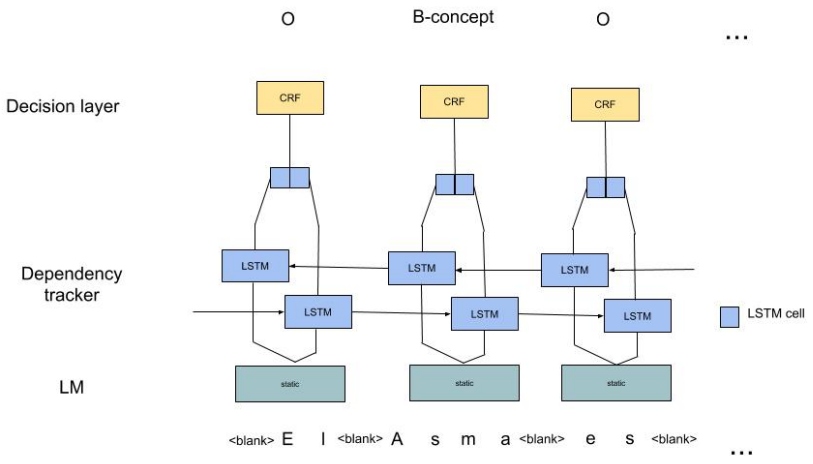

Figure 30: Architecture of Lample tagger.

5.2 Language Adaptation

We adapted XLM-RoBERTa-base tokenizer using 84771 sentences from MIMIC-III (Johnson et al., 2016) (Medical Information Mart for Intensive Care III). MIMICIII is a large, freely-available database comprising de-identified health-related data for forty thousand patients who stayed between 2001 and 2012 in critical care units of the Beth Israel Deaconess Medical Center. The database includes rich information such as vital sign measurements made at the bedside (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). We only extracted and used from the database the event-notes which correspond to the clinical notes as written in plain text by the doctor. We trained the model for 3 epoch with 8 batch size, XLM-RoBERTa-base tokenizer presented 106.66 perplexity respect to i2b2 data, after the training 11.41 perplexity was reached.

5.3 EVENT and TIMEX3 task

In this experiment 182 texts, 15424 EVENT and 2249 TIMEX3 tags were used for training and 120 texts for test, the training was only supervised by the loss function , we fine-tune the adapted XLM-RoBERTa model for EVENT and Timex3 extraction tasks. We proposed two parallel system due EVENT and TIMEX3 tags rarely overlap, both belong to clearly differentiated lexical set.

In the EVENT sequence labelling case we train the model 100 epochs with 32 batch size, in this case the fixed sequence length is 128 sub-strings. For the TIMEX3 sequence labelling case we train the model 15 epochs with 32 batch size, in this case the fixed sequence length is 150 sub-strings. In the following table 1 we can see the parameter combinations used to train each attribute prediction system.

Attribute	Epoch num	Sequence len	Batch size
EVENT type	25	32	64
EVENT modality	25	32	64
EVENT polarity	25	32	64
TIMEX3 type	25	32	64
TIMEX3 modifier	50	32	64

Table 1: Parameters table for each Attribute expert system.

In this case we run the single evaluation for Event extraction task, we achieve the relative results presented in 2. Another single evaluation for TIMEX3 extraction task achieves the results presented in 3. In EVENT extraction task 2 we show competitive performance in all prediction tasks even the main goal to achieve end-to-end performance is Span F1 metric. In TIMEX3 extraction task 3 we show competitive performance in every task except of value, this leads low overall performance regarding the primary multiplier but the main goal to achieve end-to-end performance is Span F measure metric.

Organization	Span F1	Type accuracy	Polarity accuracy	Modality accuracy	Method
Beihang University; Microsoft Research Asia, Beijing; Tsinghua	0.92	0.86	0.86	0.86	CRF
IXA research group - UPV/EHU	0.91	0.84	0.87	0.88	Transformers
Vanderbilt University	0.9	0.84	0.85	0.83	CRF+SVM
The University of Texas, Dallas	0.89	0.84	0.71	0.84	CRF+SVM
The University of Texas, Dallas—deSouza	0.88	0.8	0.76	0.05	CRF
University of Arizona, Tucson	0.88	0.71	0.8	0.8	CRF+SVM+NegEx
University of Novi Sad, Novi Sad, Serbia; University of Manchester	0.87	0.73	0.74	0.82	CRF+dictionary based
Siemens Medical Solutions	0.86	0.71	0.78	0.77	CRF+MaxEnt
MAYO Clinic	0.85	0.76	0.75	0.76	CRF
LIMSI–CNRS; INSERM; STL CNRS; LIM&BIO	0.83	0.8	0.84	0.85	CRF+SVM
University of Illinois at Urbana-Champaign	0.83	0.74	0.75	0.77	Integer Quadratic Program

Table 2: Results obtained for EVENT task in the ”i2b2 2012 challenge” competition based on (Sun et al., 2013)[table 2]

Organization	Primary score-value F-measure	Span F measure	Type accuracy	Value accuracy	Modifier accuracy	
MAYO Clinic	0.66	0.9	0.86	0.73	0.86	Regular Exp
Beihang University; Microsoft Research Asia, Beijing; Tsinghua University	0.66	0.91	0.89	0.72	0.89	CRF+SVM+rule based
University of Novi Sad, Novi Sad, Serbia; University of Manchester	0.63	0.9	0.85	0.7	0.83	Rule based
Vanderbilt University	0.61	0.87	0.85	0.7	0.85	Rule based +HeidelTime
University of Arizona, Tucson	0.61	0.88	0.81	0.69	0.8	HeidelTime +CRF
The University of Texas, Dallas	0.55	0.89	0.78	0.62	0.79	CRF+SVM+rule based
Siemens Medical Solutions	0.53	0.89	0.86	0.6	0.8	SUTime
The University of Texas, Dallas—deSouza	0.53	0.89	0.78	0.59	0.79	GUTime+CRF +rule base
Bulgarian Academy of Sciences; AMedical Sequence Labellingican University in Bulgaria; University of Colorado School of Medicine	0.49	0.8	0.72	0.61	0.71	Regular Exp
LIMSI–CNRS; INSERM; STL CNRS; LIM&BIO	0.45	0.84	0.75	0.54	0.72	HeidelTime
IXA research group - UPV/EHU	0.22	0.91	0.85	0.25	0.83	Transformers

Table 3: Results obtained for TIMEX3 task in the ”i2b2 2012 challenge” competition based on (Sun et al., 2013)[table 2]

5.4 TLINK task

For this experiment 182 texts, 15424 EVENT, 2249 TIMEX3 and 31657 TLINK tags were used for train and 120 texts for test, the training was only supervised by the loss function, we fine-tune the adapted XLM-RoBERTa model for each target prediction task. We proposed parallel systems due the different targets require different maximum sequence lengths, But we expect overlapping errors because targets share lexical set.

In the following table 4 we can see the parameter combinations used to train the each target prediction system. In this approach we didn't treat TIMEX3-EVENT and TIMEX3-TIMEX3 targets because we didn't reach a suitable model to track them correctly.

Target	Epoch num	Sequence len	Batch size	prune threshold
EVENT-SECTIME	5	300	13	INF
EVENT-EVENT	10	100	50	40
EVENT-TIMEX3	3	150	32	200

Table 4: Parameters table for each Target expert system.

We run the single evaluation for TLINK extraction task based on the EVENT / TIMEX3 ground truth, we achieve the relative results presented in 5. As we can see on table the TLINK extraction system has improvement room. The recall performance is low due the solution has problems with long dependencies and the overlapped relations.

Organization	F measure	Precision	Recall	Method
Vanderbilt University	0.69	0.71	0.67	Rule based pair selection+CRF+SVM
National Research Council Canada	0.69	0.75	0.64	MaxEnt+SVM+rule based
Beihang University; Microsoft Research Asia, Beijing; Tsinghua University	0.68	0.66	0.71	SVM
Arizona State University	0.63	0.76	0.54	SVM+rule-based
The University of Texas, Dallas—deSouza	0.61	0.54	0.72	CRF
University of California, San Diego; Department of Veterans Affairs, Tennessee Valley Healthcare System	0.59	0.65	0.54	MaxEnt/Bayes
Academia Sinica; National Taiwan University; Institute For Information Industry; Yuan Ze University	0.56	0.57	0.56	Rule based+MaxEnt
The University of Texas, Dallas	0.56	0.48	0.66	SVM
LIMSI–CNRS; INSERM; STL CNRS; LIM&BIO	0.55	0.51	0.59	SVM
Brandeis University	0.43	0.34	0.59	MaxEnt
IXA research group - UPV/EHU	0.29	0.76	0.18	Transformers

Table 5: Results obtained for TLINK task over EVENT / TIMEX3 ground truth in the "i2b2 2012 challenge" competition based on (Sun et al., 2013)[table 2]

Finally we insight the achieved results by the end-to-end system (Sun et al., 2013)[Table 4]. We ran the single evaluation for end-to-end task over the pipeline of the previously exposed EVENT and TIMEX3 task and TLINK task, we achieved the results presented in 6.

Organization	Primary score (F measure of TLINK)	F measure EVENT span	F measure Timex3 span
Vanderbilt University	0.6278	0.9011	0.8607
Beihang University ; Microsoft Research Asia, Beijing; Tsinghua University	0.5924	0.9166	0.9098
The University of Texas, Dallas	0.5258	0.8933	0.8907
The University of Texas, Dallas—deSouza	0.5126	0.8835	0.8886
LIMSI–CNRS; INSERM; STL CNRS; LIM&BIO	0.4932	0.8307	0.8385
IXA research group - UPV/EHU	0.3771	0.9132	0.9051
MAYO Clinic	0.3741	0.8548	0.8999
University of Novi Sad, Novi Sad, Serbia; University of Manchester	0.3448	0.8611	0.8607

Table 6: Results obtained for end-to-end task in the "i2b2 2012 challenge" competition based on (Sun et al., 2013)[table 4]

Finally on table 6 we can see that the framework performs competitive in end-to-end leader-board, achieves the 7Th position. As introduced before, the system is not able to track long dependencies among entities, we encounter overlapped pairs and as exposed in (Sun et al., 2013)[TLINK extraction] Among the non-section time TLINKs we encounter further detection difficulties. Finally, as the system enters in the top 8 systems, seems that even not performing well on Relation Extraction due the difficulties involved, we still remain competitive.

6 Conclusions and Future work

Processing clinical data is an emerging challenge nowadays, and its application is particularly needed in every clinical entity with large amount of patient histories such as hospitals or Medical Universities, the multilingualism is increasingly needed in our ever more global society.

In this work, we proposed a Cross-lingual approach that can automatically annotate raw clinical data with clinical events and time expressions, predict the attribute values that compose the annotations in a competitive way, we extract relations among them but results aren't competitive in this task even we show competitive in the end-to-end task, so we succeeded in our objectives: we managed to perform clinical domain adaptation with low resources available, we developed state of the art EVENT and TIMEX extraction systems, and we developed state of the art end-to-end pipeline.

The proposed solution shows competitive performance in early stage of the project, so with investment we estimate that the real deployment could be implemented as fully supported on-line service for automatic generation of clinical histories without regarding geological location, language even clinical domain. The implantation of an automatic clinical histories management system will improve the overall health care performance due to: less administrative costs may be required, less time effort for clinical professionals may be required. All those facts may lead into better clinical data management for research, diagnosis and treatment purposes.

The future work consists on: 1) improve TIMEX3 value performance using external regex based parser systems such as HeidelTime (Strötgen and Gertz, 2010), SUTIME (Chang and Manning, 2012) or TARSQI (Verhagen et al., 2005); 2) improve TLINK task results changing the entity markers distribution to ENTITY START schema, instead using '[CLS]' token we may use '[EXS]' tokens to predict relation; In addition we may study new ways to reduce the context length loosing the least information possible; 3) study how the hyperparameter combinations affect the system performance; 4) provide usefull framework as prototype.

7 Acknowledgements

I specially mention Ana Maria Santamaría Ruiz as the main promoter for the current work, her incurable problem made me think that clinical efforts and resources are extensively wasted all over the world, her mood make me feel happy and encouraged but disappointed with world, make me grow and use all my personal efforts on try to hrlp on clinical efforts.

I thank to Aitziber Atutxa Salazar, my second mom that makes me learn more than technical skills, makes me better person and brings dreams to the real world. IXA research group is my second family, they accepted me, they provided me some wings, we worked together, I met really distinguishable ones as Oscar Sainz and Oier López De LaCalle. We learnt powerful ideas and developed systems, I really appreciate them for sure we will reach whatever objective spotted. I also thank all the Teachers in the Masters their exceptional work.

I thank my colleagues, friends and team mates Mohamed Yassin, Xiao Xu, Radostina Peteva and Xaidé Caceres. Hope the shared moments may stay forever, and regarding forward to keep in touch, maybe one day we will meet again chilling outdoors the NASA, for sure they will reach whatever dream, and become society fathers / mothers. The future is today and I feel great people around.

Finally and not less important, I appreciate my close friends here in Bilbao, those I met on school, those I met on University, those I met playing video games, those I met playing chess, everyone is special we still have moments to spend together, we have room to improve our lives with good conversations and good moments.

I will keep all of you you on my heart even if we are separated by big seas or high mountains, this mention keeps forever. Cheers and keep well guys. Make me prouder over there.

Edgar Andrés Santamaría

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, August 2018a. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1139`.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018b.

Liliya Akhtyamova, Paloma Martinez, Karin Verspoor, and John Cardiff. Testing contextualized word embeddings to improve ner in spanish clinical case narratives. *BMC Medical Informatics and Decision Making*, page preprint, 02 2020. doi: 10.21203/rs.2.22697/v1.

Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzmán Edouard Grave Myle Ott Luke Zettlemoyer Alexis Conneau, Kartikay Khandelwal and Veselin Stoyanov. Xlm-roberta. `https://huggingface.co/transformers/model_doc/xlmroberta.html`, April 2020.

Edgar Andrés, Oscar Sainz, Aitziber Atutxa, and Oier Lopez de Lacalle. IXA-NER-RE at eHealth-KD Challenge 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020.

Edgar Andrés Santamaría. Identificador automático de relaciones temporales en textos clínicos basado en redes neuronales. 2019.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL `https://www.aclweb.org/anthology/P19-1279`.

Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 2012, pages 3735–3740, 2012.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019a.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. 2019b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018b.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL `https://doi.org/10.1093/bioinformatics/btz682`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and

R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Matthew E. Peters, Mark Neumann, Robert L Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *EMNLP*, 2019.

Alejandro Piad-Morffis, Yoan Gutiérrez, Hian Cañizares-Diaz, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz, and Andrés Montoyo. Overview of the ehealth knowledge discovery challenge at iberlef 2020.

Jowita Podolak and Philine Zeinert. Master thesis: Developing a cross-lingual named entity recognition model. 2020.

Sebastian Ruder. *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

Golnar Sheikhshabbafghi, Inanc Birol, and Anoop Sarkar. In-domain context-aware token embeddings improve biomedical named entity recognition. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 160–164, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5618. URL https://www.aclweb.org/anthology/W18-5618.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, Jul 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz096. URL http://dx.doi.org/10.1093/jamia/ocz096.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi Chikayama. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*, Zürich, Switzerland, September 2012.

Steven Bethard. State of the Art in Timeline Extraction. https://www.i2b2.org/events/slides/bethard-i2b2-timelines.pdf, 9. i2b2 AUG NLP Workshop.

Jannik Strötgen and Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, 2010.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154, 2014.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20 (5):806–813, 2013.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

Naushad UzZaman and James Allen. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Marc Verhagen, Inderjeet Mani, Roser Sauri, Jessica Littman, Robert Knippen, Seok Bae Jang, Anna Rumshisky, Jon Phillips, and James Pustejovsky. Automating temporal annotation with tarsqi. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 81–84, 2005.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

GuoDong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, 02 2004.