Cross-lingual Argument Mining in the Medical Domain

Author: Anar Yeginbergenova

Advisors: Rodrigo Agerri

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

June 2022

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

Laburpena

Gaur egun, domeinu medikoa gero eta arreta handiagoa jasotzen ari da Adimen Artifiziala duten aplikazioetan. Klinikariek egituratu gabeko testu-datu kopuru handi bati aurre egin behar diote pazientearen osasunari buruzko ondorio bat ateratzeko eguneroko bizitzan. Argumentu-meatzaritzak datu horiei egitura bat ematen laguntzen du, testuan argumentu-osagaiak detektatuz eta haien arteko erlazioak sailkatuz. Hala ere, Hizkuntza Naturalaren Prozesamenduan orokorrean eta testu medikoen tratamenduan bereziki lan askotan gertatzen den bezala, argumentazio konputazionalari buruzko lanaren gehiengoa ingeleserako bakarrik egin da. Hain zuzen ere, hori ere gertatzen da mediku domeinuan argudiatzeko eskuragarri dagoen datu-multzo bakarrarekin, esaterako, MEDLINE datu-baseko Randomized Controlled Trials (RCT) corpusa. Beste hizkuntzatarako etiketatutako datuen falta arintzeko, hainbat estrategia enpirikoki ikertzen ditugu testu medikoetan argudio-meatzaritza eta sailkapena egiteko. Tesi honek erakusten du ingelesetik xede-hizkuntza batera (gazteleraz) anotazioak automatikoki itzultzea eta proiektatzea modu eraginkorra dela etiketatutako datuak sortzeko eskuzko lanik egin gabe. Gainera, gure esperimentuek erakusten dute itzulpenaren eta proiekzioaren hurbilpenak gainditzen duela zero-shot hizkuntza zeharkako ikuspegiak hizkuntza-eredu eleaniztun handi bat erabiliz. Azkenik, gaztelaniaz automatikoki sortutako datuak ingelesezko jatorrizko ebaluazio ezarpenean emaitzak hobetzeko ere nola erabil daitezkeen erakusten dugu.

Abstract

Nowadays the medical domain is receiving more and more attention in the applications involving Artificial Intelligence. Clinicians have to deal with an enormous amount of unstructured textual data to make a conclusion about patient's health in their everyday life. Argument mining helps to provide a structure to such data by detecting argumentative components in the text and classifying the relations between them. However, as it is the case for many tasks in Natural Language Processing in general and in medical text processing in particular, the large majority of the work on computational argumentation has been done only for English. This is also the case with the only dataset available for argumentation in the medical domain, namely, the annotated medical data of abstracts of Randomized Controlled Trials (RCT) from the MEDLINE database. In order to mitigate the lack of annotated data for other languages, we empirically investigate several strategies to perform argument mining and classification in medical texts for a language for which no annotated data is available. This thesis shows that automatically translating and project annotations from English to a target language (Spanish) is an effective way to generate annotated data without manual intervention. Furthermore, our experiments demonstrate that the translation and projection approach outperforms zero-shot cross-lingual approaches using a large masked multilingual language model. Finally, we show how the automatically generated data in Spanish can also be used to improve results in the original English evaluation setting.

Contents

List of Figures

List of Tables

1 Introduction

Appropriate clinical decision-making is an essential part of the medical environment when the practitioner has to identify and diagnose a disease and prescribe treatment based on the patient's health condition and clinical tests. However, it can involve multiple challenges and stress, and there are many reasons for that. First, is the diversity of the symptoms, one or more of them could be a sign of multiple diseases. Second, an overwhelming amount of data from previous patients with similar symptoms. Lastly, the final decision, along with what is listed above, should take into account the latest results in the research reports. With the growth of the number of such reports and data in general, the urge of structuring information takes place, which *argument mining* tries to solve. So far, argument mining has been applied in several different domains such as law (Mochales and Ieven, 2009), biomedicine (Accuosto et al., 2021), reviews (Li et al., 2017), persuasive essays (Stab and Gurevych, 2014), with the aim of identifying argumentative structure in the data. However, the majority of those works are focused on solving the problem and creating solutions for English only. Therefore, the lack of data annotated with argumentation components is a major obstacle to work with other languages such as Spanish. In order to mitigate the lack of annotated data for other languages, this master thesis empirically investigates several cross-lingual strategies to perform argument mining and classification in medical texts for a language for which no annotated data is available.

1.1 Argument mining

Argument mining (AM) is a field of natural language processing (NLP) that focuses on extracting argumentative structures from unstructured data. By doing so, it helps to determine the notion of the view, opinion, or conclusion, and identify the proofs that either defend or oppose them depending on the context.

The main objective of argument mining is to automatically detect and define the type of argumentative components, their boundaries, and the relations holding between them.

1.1.1 Argument components

Argument components can be classified as either *Major Claim*, *Claim*, or *Premise*, and they can hold *supporting* or *attacking* links between them that create a hierarchical structure (Stab and Gurevych, 2017).

Example 1.1 below, extracted from the AbstRCT corpus Mayer et al. (2020), illustrate this. Thus, claims are marked in bold with subscript C_n, premises are in italic with subscript P_n, and major claims are in bold and italic marked as MC_n.

Claims are specific statements about the conducted experiments that hold factual information inside 1.1. It can be the conclusion from the study or the effect of the treatment. It is possible to have several claims in a single paragraph and they should be divided and treated separately.

Major Claims are a more general statements compared to a claim, and they are usually followed by a claim 1.1. In many cases, major claims are the introduction or conclusion to the claim in the corpus. However, in some tasks major claims are non-existent, and if there is one they are labeled as a claim.

Premises are the ground truth of argumentation as they report observations of studies and hold evidence for or against the claim 1.1. Therefore, a premise includes measurements and comparisons of the study. It is a continuation or description of the claim and one study can have multiple premises that either support or attack the claim.

Example 1.1 "Following pretesting in 313 patients, patients who needed district nursing and who did not need district nursing at home were randomly assigned to a control or intervention group. Intervention group patients received the Pain Education Program in the hospital, and 3 and 7 days postdischarge by telephone; this was done by nurses who were specially trained as pain counselors. Follow-up assessments were at 2, 4 and 8 weeks postdischarge. Results of the pretest showed that many patients lacked knowledge about pain and pain management. The majority of pain topics had to be discussed. [**The Pain Education Program proved to be feasible**]$_{C_1}$: *[75.0% of the patients had read the entire pain brochure, 55.7% had listened to the audio cassette, and 85.6% of pain scores were completed in the pain diary]*$_{P_1}$. [**Results showed a significant increase in pain knowledge in patients who received the Pain Education Program and a significant decrease in pain intensity**]$_{C_2}$. *[However, pain relief was mainly found in the intervention group patients without district nursing]*$_{P_2}$. [**It can be concluded that the tailored Pain Education Program is effective for cancer patients in chronic pain**]$_{C_3}$. ***[The use of the Pain Education Program by nurses should be seriously considered on oncology units]***$_{MC_1}$."

1.1.2 Argument relations

In order to build a full argumentation structure, we need to introduce relations between the argument components. Relations connect argument components to form the argumentation graphs representing the structure of an argument, where we have source and target nodes, and the edges are relation types from the source node to the target node. The links can be either attack, partial-attack, or support. There are certain restrictions on the occurrence of relations: premises can be connected to both claim or another premise, whereas claims can be connected only with another claim.

Example 1.2 *[The different schedules of vinorelbine in the two arms led to a greater survival in the NP arm without impairing the tolerance profile,]*$_{P_1}$ *[although this is not statistically significant]*$_{P_3}$. This confirms that the two-drug combination NP is a reference treatment for metastatic NSCLC. The role of three-drug combinations remains questionable in this subset of patients.

Attacking relations occur when the source component contradicts the target or when it states that some observation had no statistical significance. A attacking relation between

two arguments is represented in Example 1.2.

Partial-attacking relation is formed when a source component weakens the target but does not oppose against it. In the Example 1.1, Premise$_2$ (P_2) partially attacks Claim$_2$ (C_2), namely, it only specifies the conditions of the study without strong objection.

Supporting relations are built when the source justifies the target. In the example 1.1, there are several support links between arguments, for instance, Premise$_1$ (P_1) supports Claim$_1$ (C_1) with numerical evidences to verify the statement.

One of the major problems in many similar domain-specific tasks is that annotated data is only available for English. In order to address this issue, we will investigate several cross-lingual approaches to perform argument mining and classification in a language such as Spanish for which no labeled data exists. In order to do so, we will leverage existing labeled data in English Mayer et al. (2021) to automatically generate a Spanish version of it by using different machine translation and label projection approaches. The test data will be manually corrected to be able to experiment with a large multilingual language model in various evaluation settings: (i) a zero-shot cross-lingual approach in which we will train the model in English and evaluate it in Spanish; (ii) a translation and projection setting where we leverage the automatically generated training data for Spanish and, (iii) a multilingual evaluation in which we perform data augmentation to improve results both in the original English data and for Spanish. The generated Spanish corpus (both automatic and manually revised versions) is publicly available to encourage crosslingual research in argument mining and to facilitate reproducibility of results [1].

The main contributions of this master thesis are the following:

- We provide the first medical corpus in Spanish for argument mining by using machine translation and label projection methods;
- We perform a qualitative evaluation of the quality of the translation and projection methods for medical texts;
- We present the first experimentation on cross-lingual zero-shot and multilingual experiments for argument mining;
- We establish which strategy works best when no annotated data is available for a target language;
- We show that the automatically generated data can be used to perform data augmentation to improve results also for argument mining in the original English dataset.

We begin by reviewing existing research and state-of-the-art argument mining, previous approaches in the medical domain, and works on cross-lingual sequence labeling in Section 2.

In Section 3, we provide a comprehensive description of the source dataset that is used for the creation of the Spanish corpus, the existing argument mining pipeline, and machine

[1] `https://github.com/ragerri/antidote-projections`

translation and word alignment methods. The subsequent section Section 4 will describe the process of the corpus generation from English to Spanish. The following Section 5 regards the set of experiments to perform for argument mining. The obtained results are presented in Section 6. Discussion of analysis from the predictions and descriptions of the most frequent errors are presented in Section 7. Lastly, we provide observation and conclusion in Section 8.

2 Related work

The development of automatic argumentation started from attempts towards identifying argument structure and is closely related to theories of discourse representation described by reasoning and logic. There are several existing methods based on which aspect of the text they are focused on: Discourse Representation Theory (DRT) (Kamp et al., 2011) and Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003) are the theories that analyze the text based on linguistic characteristics of discourse. However, these theories are quite complex to be applied in practice (Bos, 2008). On the other hand, theories such as Rhetorical Structure Theory (RST) are focused more on pragmatics rather than semantic and syntactic features of the text and it has been more applied in NLP systems. Therefore, it was considered by many that RST was more approachable to represent and automatically study discourse structure and argumentation ((Peldszus and Stede, 2013; Azar, 1999; Green, 2010), etc.).

Among other works on argumentation theory, Toulmin (1958) influenced the development of argumentation by identifying different functional roles in arguments (evidence, warrant, backing, qualifier, rebuttal, and claim) based on how the conclusion is made from evidence in the text. Furthermore, Freeman (2011) investigated how to transfer them in diagramming techniques of the informal logic tradition. Dung (1995) tried to create a graph representation of argumentation based on nonmonotonic reasoning in AI and logic programming, and Peldszus and Stede (2013) introduced a diagram structure with models of textual representation of arguments and globally optimized argumentative relations. They propose that the most important relationship types in arguments are ***support*** and ***attack***. They also identified five different types of graphs based on the connections that existed between them (e.g., one claim having relations with multiple premises, claim followed by claim, etc.) shown in Figure 1.

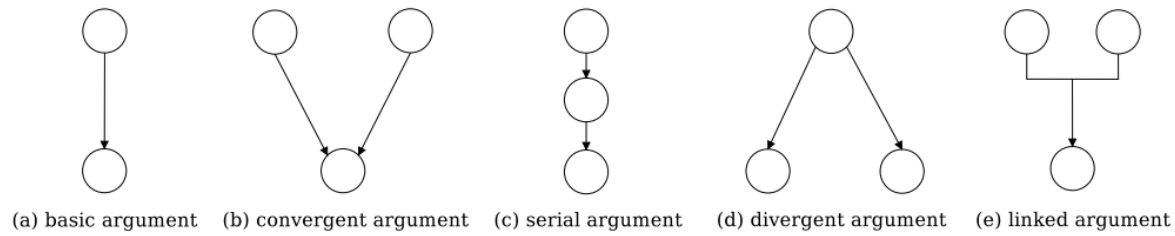

Figure 1: Argument relations diagram of an essay by Peldszus and Stede (2013)

However, Stab and Gurevych (2017) assume that the graph structure above could be somewhat ambiguous in practice when they introduced a machine learning approach for argumentation on persuasive essays. They found that the structure of the argumentative components and links in persuasive essays are somewhat hierarchical, where *Major Claim*

is the root with connections to *Claim* followed by *Premise* in the argument. The proposed structure is illustrated in Figure 2.

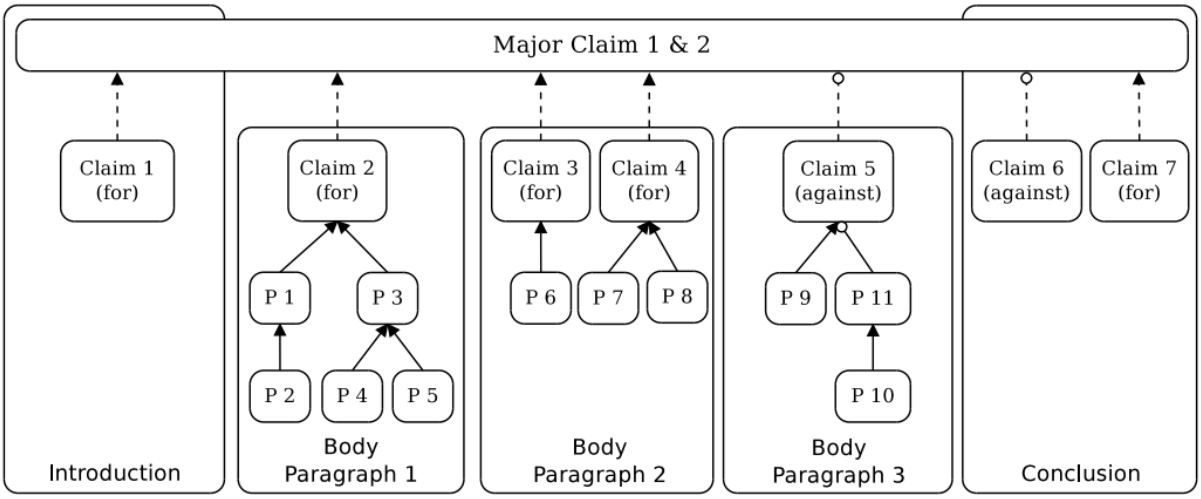

Figure 2: Argument structure of an essay by Stab and Gurevych (2017)

With theoretical knowledge of argumentation, current research interest lies on combining its theory with modern deep learning techniques. The objective here is to investigate if machine learning algorithms are able to capture an argumentative structure from given text. There are multiple works that tried to answer it using data from different domains such as education (Stab and Gurevych, 2014), law (Mochales and Ieven, 2009), news (Reed et al., 2008), science (Accuosto et al., 2021), medicine (Mayer et al., 2018), reviews (Li et al., 2017), etc. The majority of the experiments are performed using data in English and very few in other languages (Kirschner et al., 2015; Peldszus and Stede, 2015). Moens et al. (2007) introduced automatic argument detection in legal texts using general statistical features from data, a Multinomial naive Bayes classifier, and a Maximum entropy model. They reached the prediction accuracy of ∼68%. Goudas et al. (2014) tried to classify argumentative and non-argumentative sentences, and segment the arguments with Conditional Random Fields (CRF). The accuracy for distinguishing argumentative sentences was 77%. Kwon et al. (2007) focused only on identifying *Claims* and the relations they represent. To achieve it, they used a boosting algorithm and reached a F1-score of 55% for claim detection and 67% for relation classification. Stab and Gurevych (2017) created a corpus of persuasive essays in German and introduced an architecture to identify argument components and relations using SVM which is divided into 5 subtasks:

1. Identifying argument components - find arguments and set their boundaries.
2. Classification of argument components - label arguments with either major claim, claim, or premise.

3. Identification of argument relations - classify if two arguments are linked or not.

4. Generate tree - build a tree representation from the previous steps for each paragraph.

5. Stance recognition - determine any support or attack relations between arguments.

They indicated that in persuasive essays *major claim* is the root node of the argument and represents the author's standpoint. Furthermore, they consider that they are often mentioned in the introduction and conclusion. The individual paragraphs of the essays hold the actual arguments and they either *support* or *attack* author's major claim.

2.1 Argument mining in medical domain

Argument mining can be very valuable in the clinical area, particularly for experts in analyzing the impacts and results of the treatments from different sources of data. However, there are very few approaches to argument mining in the medical domain and, moreover, they all focus on solving the problem using data in English. As a consequence, there are no previous attempts to perform argument mining for the medical domain in other languages. This is partially due to the inherent difficulty if obtaining medical data to start with, but also because of the cost and complexity of obtaining the required annotations. Thus, many existing approaches solely focus on either determining argument components or only classifying argument relations in order to present structured output from unstructured medical data to assist users in decision-making.

For instance, Green et al. (2014) provided an analysis of arguments in biomedical data and created argumentation schemes and inter-argument relationships. Alamri and Stevenson (2016) created a corpus using research abstracts of studies considered in systematic reviews related to cardiovascular diseases where the objective was solely to identify contradictory claims, hence, no other information is provided in the data. Mayer et al. (2018) annotated a dataset of 169 medical abstracts and created a system to identify claims and premises in the text. Noor et al. (2017) analyzed arguments of medical drug effects following graph structure of Dung (1995). Their motivation was to identify and extract the effects of drugs from reviews on the web following argument-based analysis. Similarly, Shankar et al. (2006) described a tool for health care where part of the system deals with extracting evidence for any treatment-related claims based on Toulmin (1958)'s argumentative structure that was mentioned before. Craven et al. (2012) described the application of assumption-based argumentation to a domain of medical knowledge derived from clinical trials of drugs for breast cancer using variant-based parallel programming technique.

2.2 Cross-lingual sequence labeling

Advances in deep learning and NLP opened the gates to a world of multilinguality that allows to leverage knowledge across different languages. The idea behind cross-lingual sequence labeling is to transfer labels in-hand from annotated data in one language to

data in another language. The approach of cross-lingual sequence labeling proves to be effective when no annotated data is available in the desired language.

There are many approaches proposed for cross-lingual sequence tagging and many of them are focused on dealing with part-of-speech (POS) tagging,named-entity-recognition (NER) (Gaddy et al., 2016; Yang et al., 2017; Agerri et al., 2018; Chen et al., 2018; Liu et al., 2020), opinion target extraction (OTE) (Agerri and Rigau, 2019) and more. The majority of the approaches in transferring labels between languages require a huge amount of parallel data to create more accurate projections (David et al., 2001).

Das and Petrov (2011) introduced a bilingual graph-based unsupervised approach for the same task by building such a graph to create a connection between two languages, then projecting syntactic information to the target, and this information is used as a feature for unsupervised labels. Gaddy et al. (2016) used a coarse mapping approach to perform multilingual POS tagging and they discovered that only ten word translation pairs are enough in order to transfer POS tags effectively without the necessity in large parallel corpora. Eger et al. (2018) applied methods for label projection for AM and compared the performance of the results of the automatically translated and human translated corpus. The results showed that the performance of projection on neural machine-translated data provides results almost as good as human-translated data.

As it was noted earlier, there is no available corpus for medical argumentation in languages other than English. However, previous works that tried to solve this issue, provide us with the methods required to deal with this deficiency. Thus, in this project we will investigate the best strategy to perform argumentation in the medical domain when no data is available for a specific language using available resources, such as translation and projection and multilingual large language models such as mBERT (Devlin et al., 2019).

3 Methodology

In order to create the corpus in the target language, and apply argument mining, we first want to translate and transfer the labels of the source corpus to the target. In this section we present the medical dataset for annotated with argument mining, the machine translation and projection system used to automatically generate annotated data in Spanish and the mBERT, the language model used to perform the cross-lingual experimentation.

3.1 Data

The corpus of Randomized Clinical Trials (RCT) of medical abstracts were used to perform experiments of argument mining in this project Mayer et al. (2021). The abstracts were obtained from Evidence Based Medicine, which are clinical reports from observations of patients based on evidence for decision making. The corpus contains paragraphs of five types of diseases: neoplasm, glaucoma, diabetes, hepatitis B, and hypertension. Neoplasm was selected as a training set since it covers dysfunctionalities in the whole human body which allows for better generalization. Overall there are 500 neoplasm, 100 glaucoma, and 100 mixed (20 of each disease mentioned above) abstracts. The corpus was annotated separately for argument components and relations. The distribution of argument components in data is shown in Table 1 and the distribution of relations in Table 2.

Argument components by definition are either *Claim*, *Major Claim* or *Premise* (or *Evidence*) (Mayer et al., 2018). The number of Premises in the data is almost twice as much as Claims, while the presence of Major Claims is very reduced. This is because Claims and Major Claims are general statements or conclusions in the abstract and Premises are justifications of them which can span to several sentences in the paragraph. Claims in the corpus can be identified by the information it carries and by specific phrases, such as "According to the results", "These results support", or "This suggests"; whereas the premise mostly describes numbers and pieces of evidence from the study. Major Claims usually carry factual information about study or treatment. An example of the paragraph with all three argument components is illustrated in Figure 3 where the general statement of the paragraph is in the beginning (*Major Claim* 1 and 2) and the conclusion appears at the end of the text as *Claim*. There are two *Premises* before the last sentence that hold a supporting relation with *Claim*.

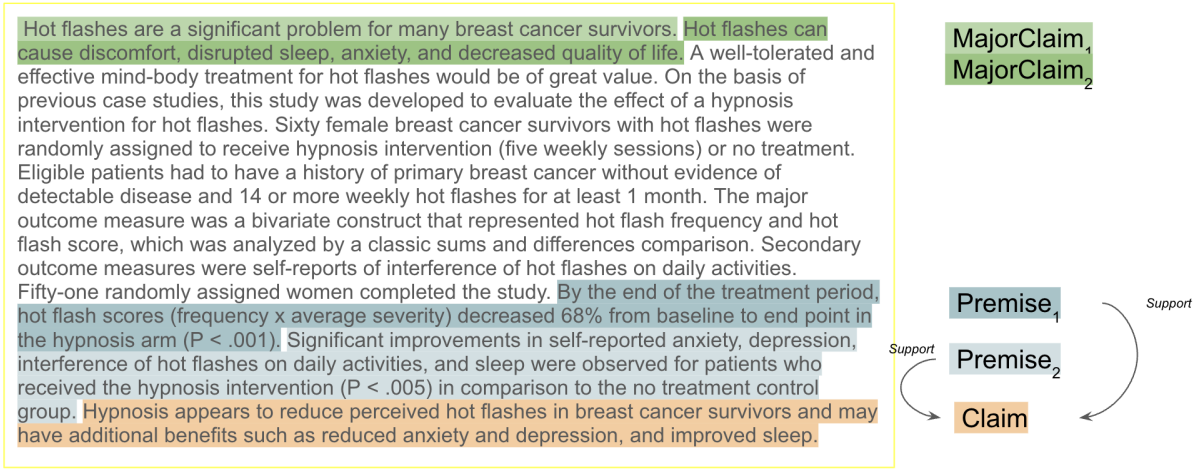

Figure 3: Example paragraph of argument relations and components. Colors denote individual argument components while the arrows refer to the relation between components.

It has been mentioned before that *Major Claims* are the stance of the paragraph and an introduction to *Claims*. Moreover, they are structurally similar to each other. As we can see on Table 1 *Major Claims* appears only in ∼3% of the data, and hence during experimental part they were merged with *Claim*. It is motivated by the fact that in random clinical trials there are no restrictions in the number of links that can form a tree, and one clinical trial can consist of several trees depending on the number of *Claims* and *Major Claims*.

Data	# of Premise	# of Claim	# of MajorClaim
Train	1537	666	64
Dev	438	228	20
Neoplasm	218	99	9
Glaucoma	404	183	7
Mixed	388	182	30
Total	2985	1358	130

Table 1: Distribution of argument components

In general, the majority of sequences in argument components are full sentences. However, sometimes depending on the context it might be longer or shorter as shown in Example 3.1. On average, sequences of Premises are longer than those of Claims or Major Claims, 23, 17, and 16 words respectively. Overall, the test set of mixed diseases has the longest amount of words per line (22) compared to the rest of the data (between 20-21).

Data	# of Support	# of Attack	# of No Relation
Train	1194	200	12892
Dev	185	30	1815
Neoplasm	359	60	3961
Glaucoma	317	29	2986
Mixed	296	24	3012

Table 2: Distribution of argument relations

Example 3.1 "...Although [**anthracyclines are effective chemotherapeutic agents for treating B-cell lymphoma ,** $_C$] adverse effects , such as bone marrow suppression and cardiotoxicity , limit their clinical application..."

Argument relations were annotated for the task of sequence classification where given two sets of argumentative sequence combinations the objective is to predict the links between them. Therefore, apart from *Support*, *Attack* and *Partial-attack*, there is an additional relation type *No Relation* that denotes that the given examples have no links. All the *Partial-attacking* relations were replaced and labeled as *Attacking* relation. Each argumentative sequence was paired with other candidate sequences and only a few of them can form a link, therefore, there is a huge imbalance between *No relation* and other classes. As shown in Example 3.2, among all the possible links there is only one **Support** relation and others have none relation between each other, which means that, in term of relations, the dataset is quite imbalanced.

Example 3.2 Argument relation sample

__**label**__**noRel** [Eight (73%) of 11 patients crossing over from 6.5 mg/m(2) per day to higher doses subsequently responded.] [The median duration of response from start of therapy could not be estimated for the 15 patients at 300 mg/m(2) per day owing to low relapse rates in 2 patients (13%); at higher doses it was 516 days.]

__**label**__**noRel** [Eight (73%) of 11 patients crossing over from 6.5 mg/m(2) per day to higher doses subsequently responded.] [The following drug-related adverse effects were reversible and treatable: hypertriglyceridemia (46 patients [79%]), hypercholesterolemia (28 patients [48%]), headache (27 patients [47%]), central hypothyroidism (23 patients [40%]), asthenia (21 patients [36%]), and leukopenia (16 patients [28%]).]

__**label**__**noRel** [Eight (73%) of 11 patients crossing over from 6.5 mg/m(2) per day to higher doses subsequently responded.] [No cases of drug-related neutropenic fever, sepsis, or death occurred.]

__**label**__**noRel** [Eight (73%) of 11 patients crossing over from 6.5 mg/m(2) per day to higher doses subsequently responded.] [Pancreatitis occurred in 3 patients with triglyceride levels higher than 14.69 mmol/L (1300 mg/dL), all of whom were taking 300 mg/m(2) or more of oral bexarotene per day.]

__***label***__***Support*** [Eight (73%) of 11 patients crossing over from 6.5 mg/m(2) per day to higher doses subsequently responded.] [Bexarotene (Targretin capsules) (the first

retinoid X receptor-selective rexinoid) was well tolerated and effective as an oral treatment for 15 (54%) of 28 patients with refractory or persistent early-stage cutaneous T-cell lymphoma at doses of 300 mg/m(2) per day.]

__**label**__**noRel** [Eight (73%) of 11 patients crossing over from 6.5 mg/m(2) per day to higher doses subsequently responded.] [Hypertriglyceridemia and hypothyroidism require monitoring but are reversible and manageable with concomitant medication.]

3.2 Transformers

The transformer is the language model that has a stacked encoder-decoder structure where the encoder turns the input sequence into a hidden representation, and the decoder turns it into the target sequence (Vaswani et al., 2017). The encoder consists N layers of two sublayers: a multi-head self-attention layer and a fully connected feed-forward neural network. It allows passing only relevant contextual information of the sequence from one encoder to another. The decoder, on the other hand, takes the output from the encoder and generates the output sequences. The structure of the decoder is similar to the encoder with an extra layer of multi-head attention to retrieve information from the output of the encoder.

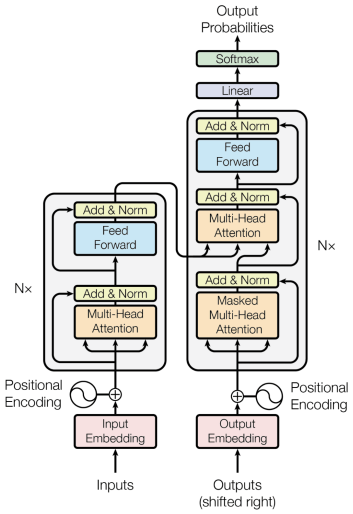

Figure 4: Transformer architecture from Vaswani et al. (2017)

Bidirectional encoder representation from Transformers (BERT) is a pre-trained transformer-based model (Devlin et al., 2018). It was trained on a huge amount of data to learn contextual word representations on two transformer tasks: 1). language modeling, where some portion of the text is masked and the model has to predict the masked token; and 2). next sentence prediction. The pre-trained model can be fine-tuned for a specific task on smaller data. BioBERT is a language model fine-tuned and pre-trained on the biomedical corpus from PubMed and outperforms BERT in the tasks related to biomedicine (Lee et al., 2019). SciBERT was pre-trained on the set of scientific papers (Beltagy et al., 2019). It also contains biomedical data, but in a smaller amount compared to BioBERT.

3.3 Argument Mining Pipeline

The argument mining pipeline consists of the following steps: 1). find the boundaries of the arguments in the text, and 2). identify the type of components, i.e. claim or premise and, 3). define the relations between components as illustrated in Figure 5.

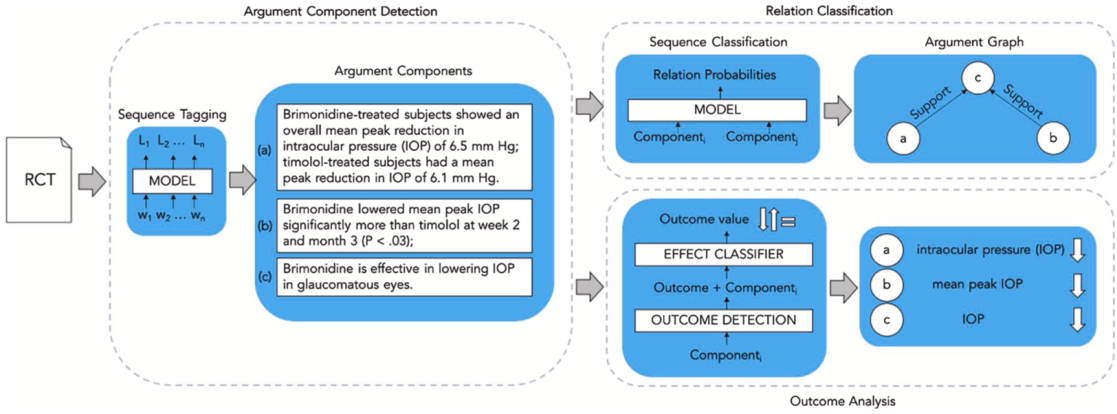

Figure 5: The full argument mining pipeline by Mayer et al. (2020).

In order to retrieve argument components from the text and identify its starting and ending boundaries, the inputs are labeled following IOB2-scheme. Mayer et al. (2020) approached this by adding the Conditional Random Fields (CRF) layer and Recurrent Neural Network (RNN), in addition to the Transformer, to the pipeline. Since the length of arguments in the dataset are considerably long (usually the whole sentence), the CRF layer helps to capture these to decide if the sequence is a part of the document by taking into account contextual information. For example, when predicting a token label only with RNNs the prediction is not dependent on the predictions of its neighbor and CRF computes transition probability that accounts for the likelihood of observing each transition between labels in the sequence. According to the authors, adding a bi-directional RNN layer, on top of the Transformer and CRF, slightly increased the prediction accuracy. They also experimented both with GRU and LSTM architectures and conclude that the former is better in identifying boundaries and differentiating between claims and premises. However, by altering parameters during training with LSTM, produced results are as good as GRU for argument component classification. Regarding the language models based on Transformers, the model performs best with SciBERT and BioBERT which are trained using the scientific and biomedical data respectively. Results obtained using the pipeline are presented in Table 3.

Relation classification, then, predicts if there is a connection between the components in the paragraph. This task is treated as a sequence classification problem where the goal is to predict relations by classifying all possible argumentative component combinations. The model is provided with two argument pairs and the objective is to predict links between

Model	Neoplasm			Glaucoma			Mixed		
	F1	F1-C	F1-P	F1	F1-C	F1-P	F1	F1-C	F1-P
SciBERT+GRU+CRF	82.41	75.84	91.11	83.97	82.89	91.68	82.40	78.21	91.35
BioBERT+GRU+CRF	80.85	73.99	90.59	83.95	83.52	91.72	82.41	78.26	92.02
BERT+GRU+CRF	82.68	76.23	89.9	82.22	79.07	89.07	82.68	77.98	89.61
SciBert+LSTM+CRF	81.99	75.58	91.23	83.06	81.87	91.76	81.93	77.23	91.52

Table 3: Results of argument component detection using different pre-trained models.

them. The links can be either *Support*, *Attack* or *No relation*. The results are provided in Table 4.

Model	Neoplasm	Glaucoma	Mixed
BERT	66.97	57.04	69.32
SciBERT	70.31	65.75	71.31
BioBERT	55.84	59.23	56.17

Table 4: F-1 score of the different pre-trained models for relation classification.

3.4 Machine Translation

The main problem in Argument Mining (AM), as well as in many natural language processing tasks, is the absence of high-quality data in languages other than English. Therefore, a possible solution is to create a corpus in the language of choice by translating the original data either by human or automatically using machine translation. The former method is considered to be more reliable. However, it requires a lot of time-consuming resources, and the latter provides fast automatic translation results which might not be accurate. Thus, in this project, we first translated a small amount of sentences using different machine translation systems which then were evaluated by two native speakers in order to identify the best system that would be able to handle the translation of the clinical data. Furthermore, four systems for automatic translation were chosen to build the desired corpus.

The corpus was translated by several freely available machine translation systems, namely, m2m-100 (Fan et al., 2021), mBART (Tang et al., 2020), OPUS-MT (Tiedemann et al., 2020) and DeepL [2]. Throughout the evaluation of the performance of each of them, the latter two systems were decided to handle the task more adequately than the others. The main issues in the domain-specific translation are specific tokens that contain numerous technical terminologies and abbreviations and are challenging for an automatic tool. The main criteria in the selection of the most accurate machine translation system, besides coherence and cohesion, was dealing with the aforementioned issues.

[2] https://www.deepl.com/

3.5 Word alignment

Word alignment is a method that is well-known for machine translation and nowadays is widely used for annotation projection. It is used as a step to transfer labels of gold annotated data to its translation. There are several existing methods for word alignments. Before the emergence of complex neural network algorithms, existing alignment methods were based on statistical approaches such as GIZA++ (Och and Ney, 2003) and Fast align (Dyer et al., 2013). These methods, or their modifications, are still considered to be reliable for different tasks requiring word alignment. With the evolution of deep learning approaches, new methods for word alignment have been developed and were used for the tag projection part of this master thesis. Those are well-known aligners for cross-lingual sequence tagging are *SimAlign* (Sabet et al., 2020) and *Awesome align* (Dou and Neubig, 2021).

Nowadays, an effective way of learning representations of text is by learning contextual word embeddings trained from multilingual language models. The Awesome aligner applies this solution by using pre-trained language models and fine-tuning them using parallel corpora to increase quality of the alignments. Since it learns word embeddings from parallel corpora, providing domain specific corpus might increase the performance of the alignments as well. On the other hand, having huge parallel corpora may not be a case for many languages and hence, the solution is to transfer word alignments based on their similarities with respect to target languages.

In our project we use the word alignment software for cross-lingual annotation projection [3] of the Abstrct English dataset to Spanish. The process and results are discussed in Section 4.

[3] `https://github.com/ikergarcia1996/Cross-lingual-Annotation-Projection`

4 Translation and Projection of Arguments

This section will describe the process of creating a Spanish corpus from the English data that was mentioned in Section 3 by translating and projecting the original English dataset. The steps of corpus creation is shown in Figure 6. First, the corpus is translated using selected machine translation systems, and then we project the BIO tags from the original to translated data. After each action, the quality of the automated output was manually inspected and corrected.

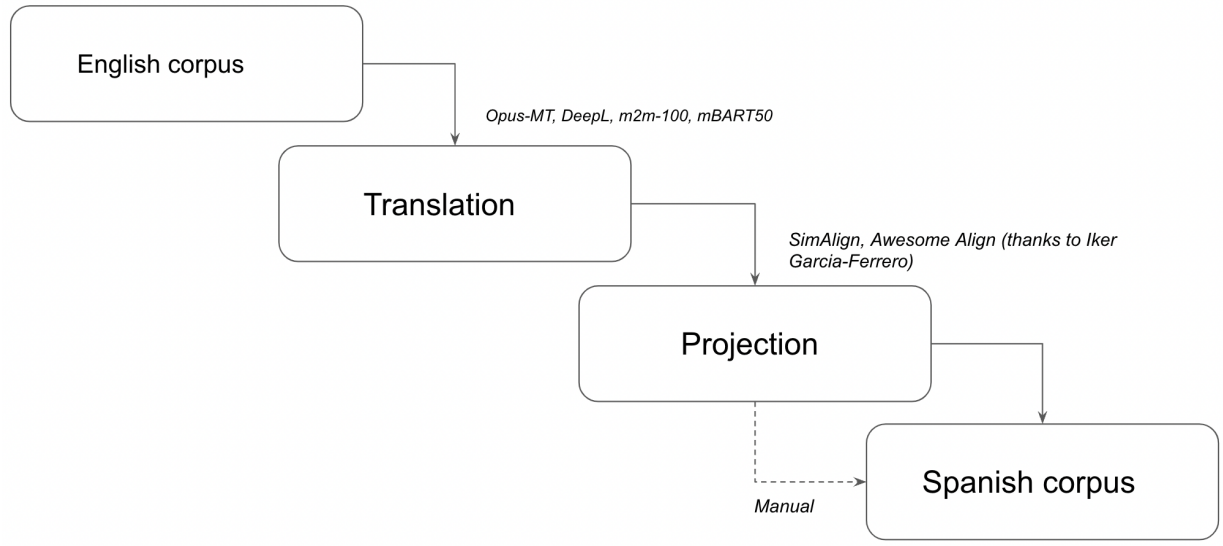

Figure 6: The process of creating Spanish data from English.

4.1 Translation

Among all the available translation systems we chose 4 widely used different neural machine translation (NMT) systems: m2m-100 (Fan et al., 2021), mBART (Tang et al., 2020), OPUS-MT (Tiedemann et al., 2020) and DeepL [4], and translated a small proportion (∼ 100 sentences) from the corpus to determine the system that generates the most accurate translations from English to Spanish and then to translate the whole corpus on the chosen NMT system. The process of selecting the best system was rather straightforward: all 100 translations in the spreadsheet where each column corresponds to one of the systems without the name were given to two native speakers to evaluate the output and choose the most appropriate ones. The inter-annotator agreement was around ∼70%. DeepL was chosen as the best performing system and OPUS-MT as the second-best performing. mBART was agreed to be the worst. Besides the quality of translations, mBART was generating random incoherent sequences in places where it was not supposed to be, hence

[4] https://www.deepl.com/

it was ruled out immediately. With respect to m2m-100, its main problem was that some expressions were not translated all, so we also discarded it.

The number of sentences translated with DeepL and OPUS-MT was the following: in the neoplasm set, 4405 for training, 679 for development and 1251 for testing. With respect to the glaucoma and mixed datasets, 1247 and 1148 sentences were respectively translated.

Example 4.1 Translation sample from the neoplasm train set:

[**EN:**] PIDs were 0·9 and 0·3 in the oxycodone/paracetamol and placebo groups respectively , on day 1 (P < 0.001) , and 1.5 and 0.3 respectively on day 3 (P < 0.001).

[$\mathbf{ES}_{\text{DeepL}}$:] Las EPIs fueron de 0-9 y 0-3 en los grupos de oxicodona/paracetamol y placebo respectivamente , en el día 1 (P < 0-001) , y de 1-5 y 0-3 respectivamente en el día 3 (P < 0-001) .

[$\mathbf{ES}_{\text{OPUS-MT}}$:] Los PID fueron 0·9 y 0·3 en los grupos de oxicodona/paracetamol y placebo respectivamente, el día 1 (P < 0·001) y 1·5 y 0·3, respectivamente, el día 3 (P < 0·001).

Overall, the quality of translations from DeepL was better than from OPUS-MT. One of the most widespread and repetitive errors in the translations was assigning wrong articles in Spanish, translation or non-translation of abbreviation, and domain-specific words. In Example 4.1, there is an acronym "PID" in the original sentence, which was translated by DeepL as "EPI" and "PID" with OPUS-MT. However, it is difficult even for humans to translate such terms without knowing the context or the full phrase of the shortened words. Here, "PID" means Pain Intensity Difference. However, it also could be Pelvic Inflammatory Disease which might correspond to the translation of the given phrase in DeepL to "EPI". However, if that was the case, then the correct translation should had been "EIP". Therefore, both systems translated it wrongly. Other issues with acronyms is illustrated by Example 4.2, where it can be seen that the abbreviated phrase is given alongside the acronym itself. However, while both systems translated the phrase equally, they provide different acronyms which do not correspond in any case with the phrase they allegedly abbreviate.

Example 4.2 Translation sample from the neoplasm train set:

[**EN**]: The primary endpoint was the Pain Intensity Difference (PID) .

[$\mathbf{ES}_{\text{DeepL}}$]: El criterio de valoración principal fue la diferencia de intensidad del dolor (DIP).

[$\mathbf{ES}_{\text{OPUS-MT}}$]: La variable principal de valoración fue la Diferencia de Intensidad del Dolor (IDP).

By analyzing the general performance of the translations, it was noted that OPUS-MT committed more mistakes in translation compared to DeepL. After deciding on DeepL to translate the Argument component data, the next step is to project the argument component labels from the original English annotated data to the automatically generated Spanish data.

Due to the available quota in DeepL, sentences for argument relation were translated using OPUS-MT only. In total, 14285 sentence pairs from the train set, 4380 sentence pairs from the test set and 2030 sentence pairs from the development set were translated.

4.2 Projection

To project the data we use two word alignment tools, namely, SimAlign (Sabet et al., 2020) and Awesome align (Dou and Neubig, 2021). SimAlign allows projecting labels without any parallel data by extracting alignments from similarity matrices of multilingual embeddings. In contrast, Awesome align requires parallel data in order to learn word embeddings, which in our case requires a parallel corpus of domain-specific data in Spanish for better results. To learn the embeddings the English-Spanish parallel biomedical corpus was provided to the model during training [5]. As a result, the overall output of both systems was considerably good, with some constant misalignment of articles in Spanish by Awesome align and detecting wrong boundaries of the label sequence in the sentences by SimAlign, and conjunctive words by both aligners. Later, these mistakes were corrected both automatically and manually.

Before running the automatic projection systems some issues in the annotation of the original corpus were identified. Namely, there were sequences with extra spaces after punctuation that during all the prepossessing steps were treated as line-breakers, i.e. creating new sentences in argument components. It caused a problem because even though it was separated as a new line, the tags still remained the same. In other words, originally the sequence was supposed to be a whole sentence labeled with BIO tags, and because of an extra space it was treated as a line-breaker, the tags were preserved when following the instructions, but the new sequence should start with 'B-' and instead it was 'I' which is not allowed since the new sequence means a new component, and hence should have starting identifier (B-).

Such sequences were identified in 12 in training, 7 in neoplasm, 4 in glaucoma, and 1 in the mixed test set lines. They were fixed manually by finding the location of the problematic token and tag in the corpus. However, by manual inspection of the corpus, more such sequences were seen in the sequences that are not part of the components (O) and they were impossible to identify. After correcting encountered issues, the projection was used over translated data. First, tags were projected automatically, then corrected semi-automatically and manually. Figure 7 illustrates the steps taken after the automatic projection of argument components.

[5] `https://github.com/biomedical-translation-corpora/corpora`

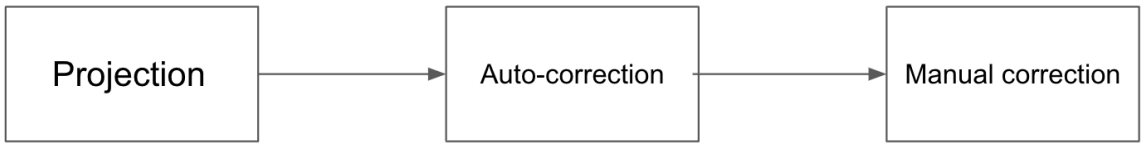

Figure 7: Annotation projection steps

Each version of the data is referred as the following: initial projection without any correction as automatic projection, the projection that was corrected programmatically as semi-automatic or post-processing, and manually corrected projection as manual projection.

4.2.1 Automatic projection

The outputs from both systems were comparably good with some repeating errors in the projection, such as missing projection of articles, conjunctive adverbs, or the wrong span of the projection. Those issues occurred the most when dealing with the sentences that are full components in the source and they were not projected accordingly in the output result. An example of the projection is shown below. In Figure 8, the source sentence with argument components is shown in green, the tokens without color are not part of the argument and have the label 'O'. The first token is labeled as "B-Premise" and the rest of the green parts before 'O' are "I-Premise", then after the uncolored part, the new argument component sequence and dot at the end, are not included in the argument component.

DC-treated patients experienced improved pain relief compared with VC (P=0.033) , whereas pain relief with DCb and VC was similar .

Figure 8: Source sentence with outlined argument component (Premise)

Projections produced by Awesome align repetitively did not align the articles as well as, although less frequently, conjunctions of different lengths in each language. For example, the misplacement of tags occurred between projections of the words "therefore" and "por lo tanto". Sometimes one-to-many and many-to-one alignments were difficult for the algorithm to find edges in the document. As illustrated in Figure 9 the majority of the argument components were transferred correctly, even conjunctive phrase was tagged correctly. However, as was mentioned above there are two articles and both of them were labeled as non-argumentative components. Moreover, the punctuation at the end of the sentence became part of the component when it is not supposed to be so.

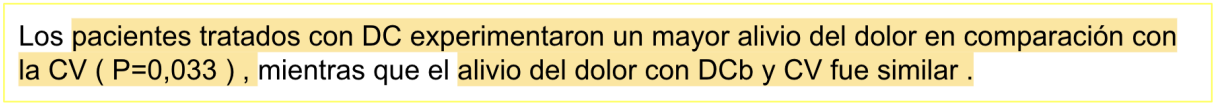

Figure 9: Projection of the sentence (Figure 8) with Awesome align

Compared to Awesome align, SimAlign handled articles relatively better, however, there is no definite pattern in alignment or misalignment of tokens. The output of the same sentence by SimAlign is shown in the Example 10. The aligner correctly projected the first article, but not the second one before "mientras que", the closing bracket and coma were also incorrectly projected. However, the dot at the end was correctly outside of the component.

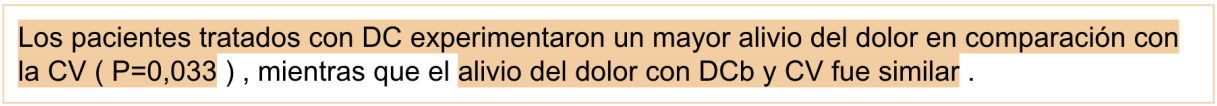

Figure 10: Projection of the sentence (Figure 8) with SimAlign

Although the results were not perfect they are still good enough and identifiable misalignment patterns can be corrected given appropriate instructions. All the issues and examples discussed above are from projection results without any modifications.

4.2.2 Post-processing and Manual projection

Since some errors from the projection of annotations follow some patterns, it is possible to correct them in some way to improve the quality of the new corpus. Therefore, after investigating the results and identifying frequent mistakes we corrected them automatically. Moreover, those that were not so easy to capture were manually corrected.

The most common issue with articles in Spanish included the sentences that are full components. Following this, we extended the projection in sentences that are full components following the logic: "if a given sequence is a full component in the source sentence then it is a full component in the translated sentence too, no matter the projection output". This allowed to decrease the number of sequences with incorrect annotation. The amount of corrected sequences in data following post-processing mentioned above is shown in Tables 5 and 6.

With respect to the training and development sets, there are overall 4405 sentences, of which 2345 are not argument components, and 1752 are sentences that are full argument components. 800 corrections in the training corpus were done in projections by Awesome align, which means that it did not align an article and/or punctuation 800 times out of 1752 where it was supposed to be full component sentences. SimAlign committed significantly

	train_awesome	train_simalign	dev_awesome	dev_simalign
overall	4405	4405	680	680
# of full O's	2345	2345	377	377
# of full component	1752	703	257	257
# of auto-corrections	800	88	95	11
# of manual-corrections	140	194	20	25

Table 5: Number of post-processed sequences, full sentence components, none-components and corrected sentences in the train and development from neoplasm translated with DeepL.

	neoplasm_a	neoplasm_s	glaucoma_a	glaucoma_s	mixed_a	mixed_s
overall	1252	1252	1248	1248	1147	1147
# of full O's	630	630	692	682	591	591
# of full component	518	518	498	506	476	480
# of auto-corrections	242	92	167	51	203	90
# of manual-corrections	51	26	26	14	47	26

Table 6: Number of post-processed sequences, full sentence components, none-components and corrected sentences in the test data translated with DeepL. _a corresponds to the results from Awesome align and _s from SimAlign.

fewer similar errors leading to the conclusion that SimAlign performed well in projecting full component sentences. The same behaviour can be observed for the three test sets, as illustrated in Table 6.

Having post-processed data still does not mean that the corpus in Spanish is correctly projected and annotated. After expanding the labels and processing the full component sequences, we looked into the rest of the projection results to correct any possible misalignments. During the manual annotation, all of the punctuation in argumentative sequences were counted as arguments, even if it was not a case in the original English corpus.

The amount of manually corrected sentences is illustrated in the last row of Tables 5 and 6. Here the overall number of corrections after post-processing is considerably lower and the difference between Awesome align and SimAlign is less than of the post-processed steps.

5 Experimental Setup

The corpus creation in Spanish was done by using the translation and projection techniques described in Section 4. First, the source corpus was translated into Spanish using OPUS-MT and DeepL and then, for each translation tag the projection is performed with both Awesome align and SimAlign. As a result, we have 4 versions of data in Spanish where they differ by translation and projection systems. Therefore, all the experiments are done for all these combinations of the systems used in Spanish, namely, OPUS-MT+Awesome, OPUS-MT+SimAlign, DeepL+Awesome, DeepL+Simalign.

The set of experiments performed for argument mining, after all the steps from previous sections and obtaining corpus in Spanish, include: zero-shot cross-lingual experiments, meaning training in English and testing on Spanish data; mixing English and Spanish corpus to train and testing individually on Spanish and English data, to see if the prediction accuracy will increase or decrease by using the automatically generated Spanish data for data augmentation; and training and testing the model using the Spanish corpus generated by translation and projection. All of the mentioned experiments are applied for both argument components detection and argument relations classification. The set of described methods are illustrated in Table 7

Experiment type	Train and development sets	Test set
zero-shot	English (mBERT)	Spanish
multilingual	English + Spanish (mBERT)	Spanish
multilingual	English + Spanish (mBERT)	English
train+project	Spanish (mBERT)	Spanish
train+project	Spanish (BETO)	Spanish

Table 7: List of experiments. The rows represent the language, and inside the parentheses the model that is used during fine-tuning.

Results obtained from in the original work using the English gold data were reported in Section 3 (Table 8). As it was mentioned before, the architecture of the model used for argument component classification is a combination of pre-trained BERT models (SciBERT, BioBERT, multilingual BERT), RNNs (GRU and LSTM), and CRF layers. To use this setting for all experiments mentioned above, Spanish multilingual BERT(mBERT) and Spanish BERT (BETO (Cañete et al., 2020)) were used in order to work with the embeddings adapted to the language. In the original work, the model was trained on 3 epochs with a learning rate of 2e-5 and batch size of 32, but during the process of hyper-parameter tuning it was determined that the performance of the model is influenced the most by learning rate hence, yielding better results when altering the learning rate to 5e-5.

Zero-shot cross-lingual argument mining assumes training the model with English corpus and testing on Spanish corpus for which we fine-tune the mBERT multilingual masked language model Devlin et al. (2019). This is also the case for the multilingual setting,

where we combine data in both English and Spanish. For experimenting with the Spanish generated data only, (train+project settings), we also included a monolingual Spanish model to evaluate its performance with respect to mBERT.

Regarding argument relations, it is a text classification task, and the classification is based on identifying relations between two sequences. A similar set of experiments are to be done for the classification of relations. However, it should be reminded that for relation classification no projection of annotations is needed, which makes the task much more straightforward.

6 Results

In this section we present the experimental results for the evaluation setup described in the previous section. After translation and projection as explained in Section 4, we have 4 versions of the corpus in Spanish from 2 translation systems and 2 projection methods for automatically projected, post-processed and manually projected data. Each version is referenced as a combination of each translation and projection used throughout this section.

6.1 English baseline

We adapted the argument pipeline originally developed for the English Abstrct dataset Mayer et al. (2021), and presented in Section 3, in order to be able to use a multilingual language model such as mBERT. Having done so, we evaluated mBERT on the English data to obtain a baseline of mBERT on this benchmark. Furthermore, we also fine-tuned the best models on this dataset according to Mayer et al. (2021). We report the results in Table 8. F1 score is an average of F1-Claim (F1-C) and F1-Premise (F1-P). The models were fine-tuned with the following hyper-parameters: 32 batch size, 3 epochs and 5e-5 learning rate for argument component and argument relation classification parts.

Argument components classification. Despite using for the experiments their pipeline and the BERT language model and their variations (BioBERT (Lee et al., 2019) and SciBERT (Beltagy et al., 2019)), we did not manage to reproduce their published results (Mayer et al., 2020, 2021). Therefore, Table 8 reports the results we obtained training those models and mBERT. BioBERT is a language model trained on data extracted from biomedical text while SciBERT was pre-trained on scientific text. The training and development corpus consists of data from neoplasm while the three test sets include data from *neoplasm*, *glaucoma* and *mixed diseases*.

Model	Neoplasm			Glaucoma			Mixed		
	F1	F1-C	F1-P	F1	F1-C	F1-P	F1	F1-C	F1-P
SciBERT+GRU+CRF	82.41	75.84	91.11	83.97	82.89	91.68	82.40	78.21	91.35
BioBERT+GRU+CRF	80.85	73.99	90.59	83.95	83.52	91.72	82.41	78.26	92.02
BERT+GRU+CRF	82.68	76.23	89.90	82.22	79.07	89.07	82.68	77.98	89.61
mBERT+GRU+CRF	82.36	74.89	89.07	80.52	75.22	84.86	81.69	75.06	88.57
SciBert+LSTM+C RF	81.99	75.58	91.23	83.06	81.87	91.76	81.93	77.23	91.52

Table 8: Results of argument component detection of the source English data. F1 is an average of F1-Claim and F1-Premise, F1-C stands for F1-Claim and F1-P for F1-Premise

From Table 8 it is obvious that the models handle identification of *Premises* considerably well compared to *Claims*. It might be because of the difference in the nature and content of those components. Premises, normally, consist of numbers and evidence, when Claims are mainly factual statements or general information.

F1-scores for neoplasm show that mBERT obtains very good results compared to specialized monolingual models such as SciBERT and BioBERT. However, mBERT results are slightly worse when evaluated out-of-domain in the glaucoma and mixed test data. In any case, the baseline shows that mBERT is competitive in this benchmark so it is a good candidate to perform cross-lingual and multilingual experiments in the following sections.

Argument relation classification. The results from the classification of argument components are not as high as the results for component detection. However, among all of the pre-trained models, the results obtained from BioBERT are the worst. On the other hand, F1-scores produced by SciBERT are the highest. All the models, other than BioBERT, perform worse on *Glaucoma* test set. Results by mBERT are worse by roughly 2 points than the original BERT but still better than BioBERT. The results obtained for relation classification are shown in Table 9.

Model	Neoplasm	Glaucoma	Mixed
BERT	66.97	57.04	69.32
SciBERT	70.31	65.75	71.31
BioBERT	55.84	59.23	56.17
mBERT	65.71	59.92	67.88

Table 9: F-1 score of the different for argument relation classification.

6.2 Experiments on Spanish projected and translated corpus

Having a model trained on English corpus our next step was to see how the model will perform when evaluated in Spanish. In this section we report the results obtained by the multilingual mBERT model when fine-tuned in English and tested in Spanish (cross-lingual zero-shot), trained and tested in Spanish (monolingual), and when trained using of both languages and testing for each language individually.

We provide results for all the versions of the Spanish corpus and it is possible to note that the performance increases with each version of the corpus from automatic to manual annotation. The same set of experiments was performed for argument relations but there is only one version of it since it only required the translation step.

6.2.1 Zero-shot results

F1-scores of zero-shot experiments for all three versions of the Spanish corpus are reported in Tables 10-12. Overall, the predictions obtained by mBERT in this zero-shot setting are surprisingly high and they get better with each improvement introduced to the corpus. Although the issue with automatic projection was that the alignment boundaries were different from the original corpus (many errors in the alignment were due to the articles in

Spanish), the prediction scores as expected were lower, compared to others, when testing on the corpus with no correction.

Model	Neoplasm			Glaucoma			Mixed		
	F1	F1-C	F1-P	F1	F1-C	F1-P	F1	F1-C	F1-P
DeepL + SimAlign	75.87	71.21	85.70	75.64	72.08	81.18	75.25	68.84	85.08
DeepL + Awesome	70.07	70.05	84.83	71.39	71.26	80.29	69.85	67.37	84.46
OPUS-MT + SimAlign	77.59	71.09	86.24	75.46	69.86	80.37	76.63	70.26	85.50
OPUS-MT + Awesome	71.36	69.74	84.97	71.22	69.70	79.59	77.98	68.94	84.70

Table 10: Zero-shot English to Spanish results of argument components with automatically projected labels

Model	Neoplasm			Glaucoma			Mixed		
	F1	F1-C	F1-P	F1	F1-C	F1-P	F1	F1-C	F1-P
DeepL + SimAlign	79.99	71.69	86.67	77.30	72.29	81.51	79.25	69.36	85.85
DeepL + Awesome	79.12	71.33	86.48	76.77	72.08	81.32	78.21	68.95	85.65
OPUS-MT + SimAlign	80.68	71.54	86.69	76.81	70.08	80.59	79.89	70.84	85.91
OPUS-MT + Awesome	80.21	71.13	86.48	76.42	69.92	80.62	79.23	70.59	85.78

Table 11: Zero-shot English to Spanish results of argument components with automatically projected labels and post-processing where labels were extended if the whole sentence is argument component

Model	Neoplasm			Glaucoma			Mixed		
	F1	F1-C	F1-P	F1	F1-C	F1-P	F1	F1-C	F1-P
DeepL + SimAlign	80.50	71.56	86.73	77.60	72.33	81.60	79.62	68.99	85.96
DeepL + Awesome	80.34	71.54	86.77	77.51	72.30	81.59	79.57	69.36	85.92
OPUS-MT + SimAlign	81.21	71.57	86.62	77.15	70.07	80.87	80.43	70.82	86.05
OPUS-MT + Awesome	81.16	71.44	86.63	77.05	70.04	80.92	80.35	71.02	85.92

Table 12: Zero-shot English to Spanish results of argument components with manual projections

The output after correcting the article issue in Spanish improved significantly by several points (Table 11). The major part of the correction took place for projections by Awesome align and, accordingly, the results of these versions of data increased considerably from 70.07% to 79.12% for DeepL+Awesome and from 71.36% to 80.21% for OPUS-MT+Awesome respectively.

Further improvements in results were reached after running the same experiments on the manually corrected data (Table 12). This time the improvement is not as significant as in the previous one although F1-score is still improved by some points. The results of DeepL+Awesome of the mixed test set improved by almost 10% when from automatically

to semi-automatically sets, and then, improved by 1% from semi-automatic to manual projection prediction. An overview of these results is provided by Table 13 where each column holds F1-macro scores obtained from zero-shot experiments for each combination of translation and projection.

Model	Neoplasm			Glaucoma			Mixed		
	Auto	Semi-auto	Manual	Auto	Semi-auto	Manual	Auto	Semi-auto	Manual
DeepL + SimAlign	75.87	79.99	80.50	75.64	77.30	77.60	75.25	79.25	79.62
DeepL + Awesome	70.07	79.12	80.34	71.39	76.77	77.51	69.85	78.21	79.57
OPUS-MT + SimAlign	77.59	80.68	81.21	75.46	76.81	77.15	76.63	79.89	80.43
OPUS-MT + Awesome	71.36	80.21	81.16	71.22	76.42	77.05	71.36	80.21	81.16

Table 13: F1 scores of each corpus version for each disease from zero-shot English to Spanish experiment

6.2.2 Train and test on translated and projected data in Spanish

The same model was trained and tested using Spanish training and testing set to observe and compare the results obtained from cross-lingual zero-shot experiments. The objective was to see if a model trained in one language with automatically created data for that language will improve the performance of the model compared to the zero-shot predictions. The model was trained by using two BERT variations: multilingual BERT, which was pre-trained on 104 different languages, and BETO, which is a type of BERT trained on a large corpus in Spanish. The model was trained on data translated with OPUS-MT and projected with Awesome align which then was manually corrected. The results of described experiments are shown in Table 14.

Model	Neoplasm			Glaucoma			Mixed		
	F1	F1-C	F1-P	F1	F1-C	F1-P	F1	F1-C	F1-P
DeepL + SimAlign (mBERT)	83.57	75.95	90.01	80.83	75.44	86.11	82.62	74.62	88.81
DeepL + SimAlign (BETO)	83.19	74.66	89.31	84.16	80.98	89.99	83.54	74.77	90.87
DeepL + Awesome (mBERT)	83.40	77.11	89.18	81.11	76.16	87.35	81.88	73.71	88.50
DeepL + Awesome (BETO)	82.84	74.70	89.57	83.78	79.95	89.93	84.15	75.89	91.11
OPUS-MT + SimAlign (mBERT)	83.03	74.68	88.69	82.06	75.76	87.36	82.64	72.94	89.31
OPUS-MT + SimAlign (BETO)	82.19	73.73	89.58	83.23	80.15	89.41	82.65	75.01	90.19
OPUS-MT + Awesome (mBERT)	82.66	74.07	88.69	82.44	76.55	87.44	82.55	73.70	89.18
OPUS-MT + Awesome (BETO)	81.91	72.96	89.56	83.30	80.69	89.45	82.48	74.03	90.36

Table 14: Train and test in Spanish with mBERT and BETO

It is noticeable that mBERT predictions have increased compared to the model trained on English and tested on Spanish corpus. Moreover, the F1-score is notably higher than

the one from the English baseline (Table 8) even when a domain-specific model was utilized. The results produced by the model with mBERT are higher compared to BETO for neoplasm, and for other test sets, it is in general lower than BETO. The improvement might be caused by a slight difference in annotations because, as it was mentioned before, the original English corpus has some inconsistencies in annotating punctuation.

6.2.3 Train and test on merged English and Spanish data

Lastly, after zero-shot and Spanish-only experiments, we merged the training sets in two languages to see if it will increase or decrease the prediction quality of monolingual test sets in English and Spanish. It can be concluded from Table 15 that this data augmentation approach increased the quality of the overall performance by 1-2% compared to training on monolingual and zero-shot results, even when evaluated in the original English data. It should also be note that the improvements are particularly large when mBERT is evaluated out-of-domain, namely, on the glaucoma and mixed test sets.

Model	Neoplasm			Glaucoma			Mixed		
	F1	F1-C	F1-P	F1	F1-C	F1-P	F1	F1-C	F1-P
Test: EN	83.51	73.42	89.38	85.31	81.05	86.73	83.63	74.98	89.25
Test: DeepL + SimAlign (ES)	84.35	76.64	88.43	84.54	78.67	87.24	83.90	73.46	88.87
Test: DeepL + Awesome (ES)	84.58	76.77	88.61	84.62	78.71	87.23	84.03	73.99	88.92
Test: OPUS-MT + SimAlign (ES)	84.54	76.02	88.89	84.20	78.69	87.31	83.18	72.15	88.98
Test: OPUS-MT + Awesome (ES)	84.35	75.76	88.89	84.25	78.74	87.26	83.16	72.46	88.84

Table 15: Train and test with merged English and Spanish dataset using mBERT

Another interesting aspect worth mentioning is that across evaluations there are no significant differences with respect to which combination of machine translation and projection model we are using. Instead, differences are given by the models and the amount of data used for fine-tuning, as our results on multilingual training demonstrate.

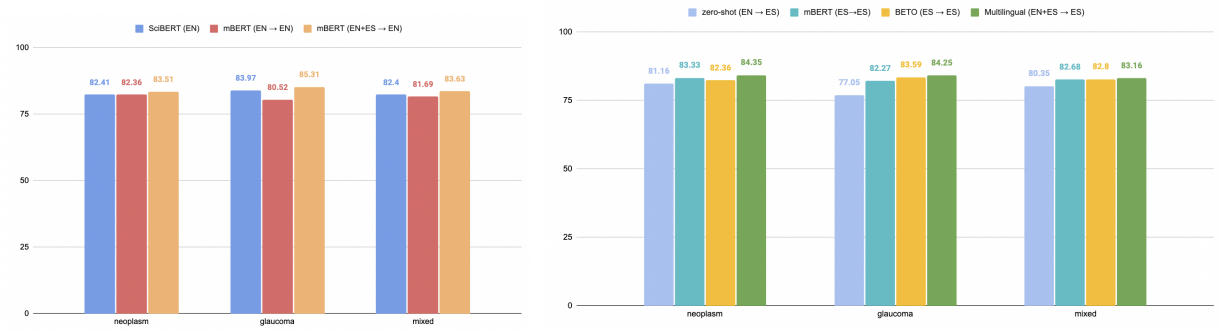

Figure 11: Results for English (left) and Spanish (right) experiments

Figure 11 provides an overview of the most important comparisons that can be done with the results of our experimentation. First, we can conclude that the best predictions are obtained from the model that was trained on the merged English and Spanish data, surpassing also the model trained with gold standard English data. This result indicates that our generated Spanish data can be used to apply data augmentation techniques. Second, the translating and projecting obtain significant better results than predicting in a zero-shot cross-lingual setting using a multilingual model such as mBERT. Third, for Spanish the monolingual model BETO performed better than mBERT, although the latter benefits from multilingual training.

6.3 Experiments with argument relation classification.

The aim of the argument relation classification is to resolve the relation types between arguments. The relation types are *Attack*, *Support* and *No Relation*. Each line in the corpus consists of two sentences, source and target, and the task for the classifier is to identify which type of relation holds between them. The source node is given several candidates as a target, and only one or two of them hold a supporting or attacking relation. All the experiments that were done for argument components detection were also applied for the classification of argument relations, with the exception that for this data no projections are necessary. Results are shown in Table 16 and in Figure 12.

Model	Neoplasm	Glaucoma	Mixed
mBERT (train: EN+ES → test: EN)	65.55	58.79	67.82
mBERT (train: EN+ES → test: ES)	62.55	58.60	65.74
mBERT (train: EN → test: ES)	62.45	55.92	65.02
mBERT (train and test: ES)	63.25	54.35	65.40
BETO (train and test: ES)	65.27	60.15	66.80

Table 16: F-1 scores of the different pre-trained models for relation classification

Here, similar to the results of previous experiments, the prediction scores are significantly lower for *Glaucoma* and are higher for *Mixed* test sets. Furthermore, in this setting the models fine-tuned with multilingual data (both Spanish and English) are not better than their counterparts trained on the target languages. Furthermore, while mBERT trained and tested on the translated and projected data is better than when applied in a zero-shot scenario, differences are not as large as for Argument Component Detection. In fact, zero-shot performs better for Glaucoma. Finally, the monolingual BETO model does obtain better results than mBERT for most of the evaluations on relation classification.

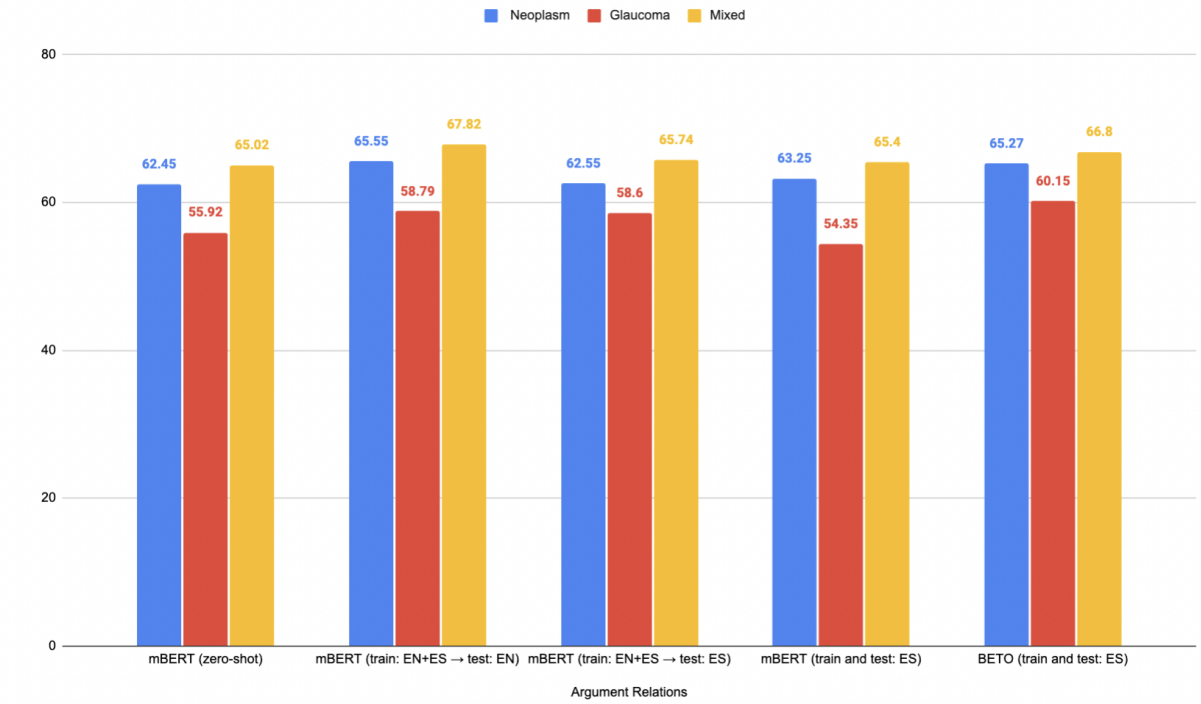

Figure 12: Relation classification results under different experimental setups

Another aspect to mention is that the relation classification corpus is extremely imbalanced (as shown in Section 3). We believe that this may be one of the reasons why the classification quality is significantly low. Another reason is the lack of context in the task itself, given that in many cases it is extremely difficult to distinguish relations given only two sentences, without further context.

Thus, we believe that relation classification in the medical domain cannot be straightforwardly determined based only on local textual information. It may require more complex structures and additional insights from the data to be able to identify argumentative structures.

7 Error Analysis

From the results obtained in previous sections, it has been clear that SciBERT with GRU layer is the best performing model for the English data set. Furthermore, mBERT performed better when trained on multilingual data. On top of that, prediction quality by the multilingual model (mBERT) was in general quite competitive good, which opened the door for other types of cross-lingual experiments. In this section we will provide a qualitative analysis of the predictions produced by the models with the aim of identifying the most important errors.

Before diving into a detailed analysis of outputs generated under each individual setting, there are some errors in the predictions throughout all the experiments that can be more or less generalized. For instance, the majority of the wrong classifications were in assigning correct IOB-tags and their boundaries, along with incorrect argument types. Overall, identification of *Premise* were more accurate compared to *Claim*. In fact, the majority of the misclassifications happened in determining *Claim* arguments. Comparisons to analyze predictions across models in this section were made on the *mixed test set*, because it has texts of all 5 diseases in it. The number of erroneous outputs is shown in Table 17.

model	# of misclassifications
SciBERT+GRU (EN → EN)	150
SciBERT+LSTM(EN → EN)	142
mBERT (EN → EN)	154
mBERT (EN → ES)	156
mBERT (ES → ES)	151
mBERT (EN+ES → ES)	158
mBERT (EN+ES → EN)	152
BETO (ES → ES)	148

Table 17: Number of erroneous predictions for mixed test set using different models. The first column is in the following form: model (train set language → test set language). The results for Spanish data are from manually refined projections.

As we can see from the table the numbers are not too scattered, we may assume that majority are the same sequences, at least among languages. The most common mistakes across all models are described below.

The majority of the tags are correct except for random tokens in different parts of the sequence with another class (Example 7.1). It is difficult to follow any pattern in this situation, nevertheless, in some cases, it assigns 'O' only to punctuation and any argument type to the other part of the text.

Example 7.1 . The sentence is *Premise* and bold tokens were classified as *I-Claim*

Text: el control de los sintomas en ambos brazos fue **similar** para los sintomas **especificos** de la enfermedad , como tos , dis ##nea , dolor o hem ##op ##tis ##is .

Another frequent mistake is when one sentence holds several arguments, but the model could only recognize one and assign it to the whole set. Similar behavior when dealing with lengthy sequences. (Example 7.2).

Example 7.2 . The first part of the sentence is *Claim* (in bold), ',' and 'y' are outside of argument and the rest are *Premise*, but model labeled everything as *Premise* (in italic).

Text: **ambos proc ##edi ##mientos prod ##uje ##ron una reduccion estadistica ##mente significativa de la pio** , y *los ojos some ##tidos a im ##ct alcanzar ##on una pio menor que los ojos del grupo de pt ##c a los 12 meses de segui ##miento (9 , 5 \hxc2 \xb1 2 , 4 mmhg y 11 , 7 \xc2 \xb1 2 , 1 mm hg , respectivamente , p < 0 , 001) .*

The model sometimes fails to identify any arguments in the input. In some examples *Claim* is classified as *Premise* and *Premise* as *Claim*. Apparently, some argumentative sequences are not explicitly identifiable compared to others.

One of the major error types is including conjunctive words and punctuation in the argument. However, it is difficult to say if those tokens should be counted as arguments in the annotated data as well. This is probaly due to inconsistencies in annotating the original English data. Lastly, in some examples, the beginning token is classified as one component and the rest of the text as another component (Example 7.3).

Example 7.3 . Here, the first token is labeled as *B-Premise* and the rest as *I-Claim*, whereas the whole sentence is supposed to be *Claim*

Text: **en** cuanto a la calidad de vida posto ##pera ##toria , los pacientes some ##tidos a qui ##mio ##tera ##pia intra ##arter ##ial parecia ##n estar en una situacion lige ##ramente mas favorable .

It was mentioned previously that we could achieve comparable results with LSTM when increasing the learning rate. Before that GRU model was outperforming it by almost 20% and by looking at the quality of the predictions, before the changes LSTM model made mistakes in 401 sentences in the mixed test set and it decreased to 142 after setting an appropriate learning rate. In general, the majority of the errors were due to the model's failure in finding correct boundaries of arguments and, in cases when it did, failing to find the beginning of the argument, i.e. 'B-' token(Example 7.4). This issue has been improved afterwards. Moreover, it outperformed some tests on the GRU model.

Example 7.4 LSTM model correctly found the argument component but not in IOB

Text: both therapies were well tolerated .
True tags: B-Claim I-Claim I-Claim I-Claim I-Claim I-Claim

Predicted tags: I-Claim I-Claim I-Claim I-Claim I-Claim I-Claim

In the case of outputs from manually projected Spanish data, using BETO and mBERT, it can be concluded that the aforementioned general errors were seen here as well. Most of the time the incorrect predictions were in the same sentences. However, there are some differences that are worth mentioning. The most noticeable one is tokenization: mBERT splits words to more atomic levels than BETO. For instance, for the word "complicaciones", mBERT tokenized it as "comp, ##lica, ##ciones" while with BETO the token remained unchanged. Another point is that even though the predictions are incorrect by both models, it is more likely that BETO recognizes argumentative sentences more frequently than mBERT. In such examples, the former either identifies the wrong argumentative component or incorrect IOB-tags while the latter tags everything as 'O'.

Prediction errors from zero-shot experiments vary between each version of projected Spanish data. Initially, most of the errors were because in the test set the articles were omitted by the projection system. In other words, the model tags the token as part of the argument component while the "true" label indicates that it is not, hence decreasing the accuracy. This issue improved after fixing the article problem in the test set. However, this time, the model would correctly predict the tags, except for some parts of the sequence for many inputs on the manually corrected corpus (Example 7.5). This issue improved when using manually projected Spanish corpus.

Example 7.5 In the following example the whole sentence is predicted as *Claim* and bold part as *Premise*. The true values are *Claim*.

la administracion de gs - 962 ##0 **durante 12 semanas** no tuvo ningun efecto significativo sobre los niveles seri ##cos de anti ##geno ##s de superficie de hepatitis b , pero pare ##cio aumentar las respuesta ##s de celulas t y celulas nk y b'reducir' b'la' capacidad de nk para sup ##rimi ##r las celulas t .

To sum up, the errors specific to zero-shot experiments, correctly predict the component but add different one in the random parts of the sequence, was improved by wrong IOB borders and incorrect classification in general. The former somewhat improved when training on the merged Spanish and English corpus. At the end, the most common misclassification occured when trying to identify the right argument component type, especially, *Claim*.

Argument relations. Since the corpus for argument relations are imbalanced the prediction results are considerably low. Nevertheless, for the clinical data, determining the relations between arguments seems to require more information than solely relying on the information from two sentences without providing the whole context. In Example 7.6, the relation type is *Support* and *No Relation* in 7.7. It is not obvious, even for humans, how this relation types are motivated without knowing the context. First, it is not obvious what kind of patients is the NGT group. Second, we are not given information about what is NGT.

Example 7.6 **__label__Support** [Patients in the NGT group reported significantly (P<0.05) better scores of QoL at both 6 months and 1 year.] [Patients who underwent gastric tube reconstruction develop less postoperative digestive tract complications, and have a quicker recovery and a better QoL during the follow-up period.]

Example 7.7 **__label__noRel** [Patients in the NGT group reported significantly (P<0.05) better scores of QoL at both 6 months and 1 year.] [Regarding the QoL investigation, the scores of QoL dropped for all patients at 3 weeks after surgery.]

To sum up, it is not straightforward to identify relation types from source sentences to the target without any context provided which may explain the difficult for the models to perform better in this particular task.

8 Concluding Remarks

In this thesis we have investigated several strategies to perform argument mining and classification on medical data for a language for which not available data is available. We have taken a real case scenario in which the only dataset annotated with argument structures for the medical domain is in English. Taking this as a starting point, we have explored two avenues to be able to perform the task in Spanish. A first option is to leverage large multilingual language models such as mBERT to perform transfer learning, namely, learning on the available English data and predict in Spanish. A second method is proposed by automatically generating data in Spanish via machine translation and label projection.

In order to create an annotated dataset in Spanish for argumentation in the medical domain, we first machine translated the source English corpus using the OPUS-MT and DeepL machine translation systems. Then the annotations were automatically projected from English to the machine translated data using word alignment tools. Next, the data in Spanish was corrected automatically and manually.

The obtained results indicate that the generated Spanish data helps to perform data augmentation which is highly beneficial to improve results for both English and Spanish benchmarks. Furthermore, experiments indicate that for this domain the translation and projection approach performs better than the zero-shot cross-lingual transfer.

The results from zero-shot experiments on Spanish data were good enough but slightly lower than when the model was trained on Spanish corpus. The main detected error from the zero-shot predictions compared to other experiments in assigning the wrong tag in the middle of the sequence

Another issue was the misclassification of punctuation and linking words. We hypothesize that this issue was well-handled when the model was trained on Spanish data because during manual correction all types of punctuation were annotated as the preceding tokens.

For relation classification the highest scores were obtained by SciBERT, then BERT, and mBERT for English data. Cross-lingual zero-shot results were lower to $sim3\%$, but monolingual Spanish and mixed English and Spanish data improved the prediction almost to the level of original English results. Overall, the predictions were consistently worse for *Glaucoma*.

Finally, apart from the scientific findings, we should stress that in this thesis have created first dataset Spanish to perform argumentation mining and classification in the medical domain. Based on this, our work constitutes the first to provide an in-depth study and empirical experimentation on cross-lingual methods for argument mining and classification.

Further work should include trying newer multilingual language models such as XLM-RoBERTa and mDeBERTa, which may help to improve results even over domain-specific English language models such as SciBERT or BioBERT. Furthermore, we would like to further explore the method presented in this thesis to experiment with computational approaches to argumentation to other specific domains and languages for which no annotated data is available.

References

Pablo Accuosto, Mariana Neves, and Horacio Saggion. Argumentation mining in scientific literature: From computational linguistics to biomedicine. In *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36.* CEUR Workshop Proceedings, 2021.

Rodrigo Agerri and German Rigau. Language independent sequence labelling for opinion target extraction. *Artificial Intelligence*, 268:85–95, 2019.

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Abdulaziz Alamri and Mark Stevenson. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of biomedical semantics*, 7(1): 1–9, 2016.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

Moshe Azar. Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation*, 13(1):97–114, 1999.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Johan Bos. Wide-coverage semantic analysis with boxer. In *Semantics in text processing. step 2008 conference proceedings*, pages 277–286, 2008.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. Multi-source cross-lingual model transfer: Learning what to share. *arXiv preprint arXiv:1810.03552*, 2018.

Robert Craven, Francesca Toni, Cristian Cadar, Adrian Hadad, and Matthew Williams. Efficient argumentation for medical decision-making. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. 2011.

Yarowsky David, Ngai Grace, Wicentowski Richard, et al. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8, 2001.

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*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*, 2021.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77 (2):321–357, 1995.

Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, 2013.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! *arXiv preprint arXiv:1807.08998*, 2018.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.

James B Freeman. *Argument Structure:: Representation and Theory*, volume 18. Springer Science & Business Media, 2011.

David M Gaddy, Yuan Zhang, Regina Barzilay, and Tommi S Jaakkola. Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. Association for Computational Linguistics, 2016.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and social media. In *Hellenic Conference on Artificial Intelligence*, pages 287–299. Springer, 2014.

Nancy Green, E Cabrio, S Villata, and A Wyner. Argumentation for scientific claims in a biomedical research article. In *ArgNLP*, pages 21–25, 2014.

Nancy L Green. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, 24(2):181–196, 2010.

Hans Kamp, Josef van Genabith, and Uwe Reyle. Discourse representation theory. In *Handbook of philosophical logic*, pages 125–394. Springer, 2011.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, 2015.

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W Shulman. Identifying and classifying subjective claims. In *DG. O*, pages 76–81. Citeseer, 2007.

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, btz682, 2019.

Mengxue Li, Shiqiang Geng, Yang Gao, Shuhua Peng, Haijing Liu, and Hao Wang. Crowdsourcing argumentation structures in chinese hotel reviews. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 87–92. IEEE, 2017.

Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. On the importance of word order information in cross-lingual sequence labeling. *arXiv preprint arXiv:2001.11164*, 2020.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. Argument mining on clinical trials. In *COMMA*, pages 137–148, 2018.

Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press, 2020.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artificial Intelligence in Medicine*, 118:102098, 2021.

Raquel Mochales and Aagje Ieven. Creating an argumentation corpus: do theories apply to real arguments? a case study on the legal argumentation of the echr. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 21–30, 2009.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230, 2007.

Kawsar Noor, Anthony Hunter, and Astrid Mayer. Analysis of medical arguments from patient experiences expressed on the social web. In *International conference on industrial, engineering and other applications of applied intelligent systems*, pages 285–294. Springer, 2017.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.

Andreas Peldszus and Manfred Stede. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, 2015.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008*, pages 2613–2618. ELRA, 2008.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*, 2020.

Ravi D Shankar, Samson W Tu, and Mark A Musen. Medical arguments in an automated health care system. In *AAAI Spring Symposium: Argumentation for Consumers of Healthcare*, pages 96–104, 2006.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, September 2017.

Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510, 2014.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.

Jörg Tiedemann, Santhosh Thottingal, et al. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 2020.

Stephen E Toulmin. *The uses of argument*. Cambridge university press, 1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*, 2017.