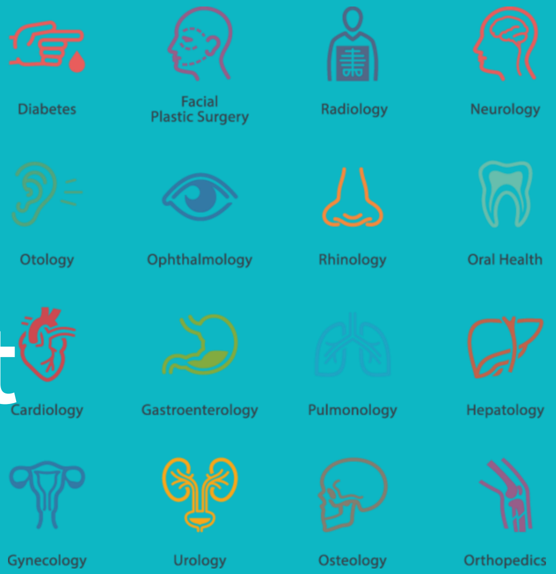
Transfer learning for classifying Spanish and English text by clinical specialties



Alexandra POMARES-QUIMBAYA Javeriana University, Bogotá, Colombia
Pilar LÓPEZ-ÚBEDA Jaén University, Spain
Stefan SCHULZ Medical University of Graz, Austria

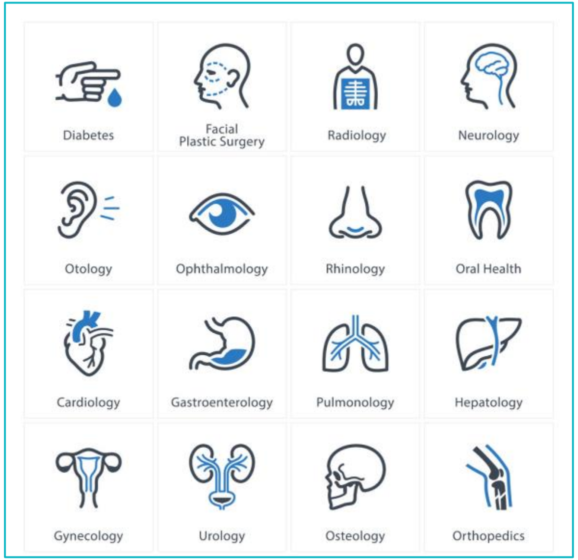
31st Medical Informatics Europe Conference - May 31, 2021

Rationale for classifying text by clinical specialty

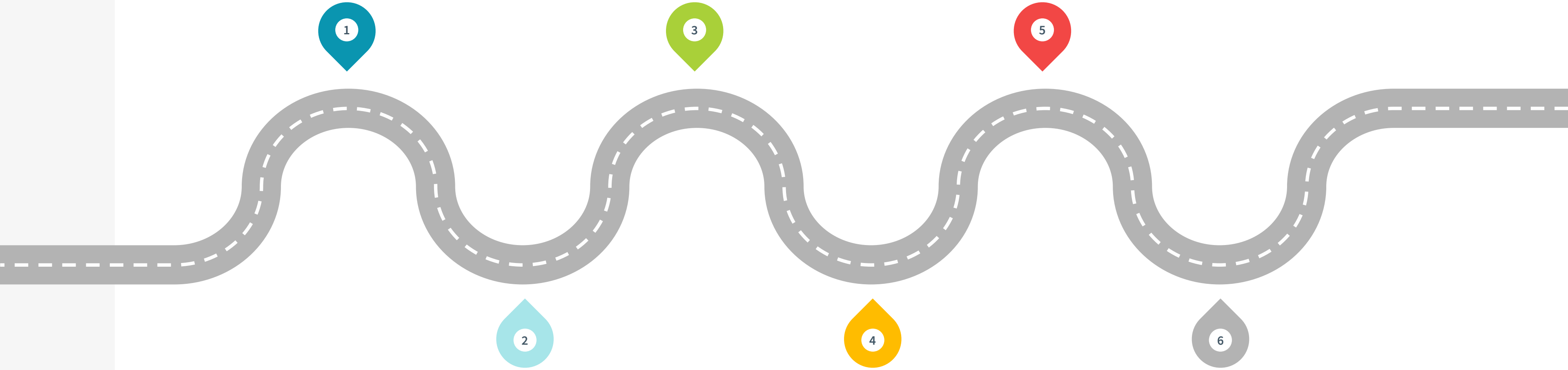
- Language varies between clinical specialties
- Knowing the specialty adds important context for better text interpretation, e.g.
  - “RTA” = “Road traffic accident” (traumatology)
    vs. “RTA” = “Renal tubular acidosis” (nephrology)
  - “body”, “infarction”, “tube”, “nail”, ….
- Train language models from open data and use them for classifying text by the specialty(ies) it belongs to?

Background

- Multi-label classification by clinical specialties
- Improving biomedical language processing
- Transfer learning using transformers
- Languages: Spanish and English



Steps



1. Selection of clinical specialties
2. Generation of the Spanish and English training dataset
3. Generation of the Spanish test dataset / corpus
4. Machine Translation of the corpus to English
5. Training of models of the BERT family
6. Evaluation

Selecting clinical specialties

Clinical Specialties

['Internal Medicine', 'Nuclear medicine', 'Radiology', 'Pediatrics', 'Dermatology & Venereology', 'Anesthesiology', 'Orthopedics & Traumatology', 'Gynecology & Obstetrics', 'Rehabilitation Medicine', 'Forensic Medicine', 'Ophthalmology', 'Neurology', 'Psychiatry', 'Urology', 'Surgery', 'Otolaryngology', 'Pathology', 'Family Medicine']

Training dataset: titles and abstracts

TRAINING DATA GENERATION

Extraction of articles from MEDLINE using queries

- Title or abstract in Spanish and English
- Manual creation of filters for each specialty
- Excluded case reports (reserved for testing)

Statistics:

- Number of records: 194,527
- Number of specialties: 18

Partitions: 80% training + 20% development

Query example for CARDIOLOGY

```
SPA[LA] AND
not Case Reports[PT] AND
("Cardiology"[TA] OR "Cardiology"[TIAB] OR
"Cardiology"[MH:noexp] OR "Cardiology"[SH:noexp] OR
"Cardiology"[PT] OR "Cardiology"[PS] OR "Cardiology"[CN]
OR "Cardiology"[SI] OR "Cardiology"[OT] OR
"Cardiology"[AD] OR "Cardiología"[TA] OR
"Cardiología"[TIAB] OR "Cardiología"[MH:noexp] OR
"Cardiología"[SH:noexp] OR "Cardiología"[PT] OR
"Cardiología"[PS] OR "Cardiología"[CN] OR
"Cardiología"[SI] OR "Cardiología"[OT] OR
"Cardiología"[AD] OR "Cardiovascular Diseases"[MH] OR
"Heart Diseases"[MH] OR "Vascular Diseases"[MH] )
```

Spanish test corpus: case descriptions from full texts

TEST DATA GENERATION

Extraction of articles from MEDLINE using queries

- Annotated with Publication Type **Case reports** (as proxies for “real” clinical documents)
- Accessible full texts
- Reports manually extracted from the full texts
- Manually annotated by medical specialty(ies)

Statistics:

- 227 articles
- 263 case descriptions

Case Description Example - Dermatology

Se trata de un niño de 15 meses de edad, previamente sano. Consulta a su pediatra de cabecera por presentar pápulas eritematosas en muñeca izquierda. Algunas de ellas se tornaron costrosas y en pocos días se sumaron pápulas, placas y pequeños nódulos eritematosos con escamas en axila derecha, región supraumbilical y axila izquierda. Se encontraba en buen estado general y afebril.

Machine translation of Spanish case descriptions to English

GENERATION OF A PARALLEL SPANISH-ENGLISH CORPUS

Library: Google Translate

Spanish case description example

Un paciente masculino de 30 años de edad, 10 años de evolución previo a la cirugía, con lesión aislada longitudinal en el cuerno posterior del menisco medial, zona roja-blanca, sin asociación con lesión ligamentaria reparado con dos suturas y dos Fastener.

English case description example

A 30-year-old male patient, 10 years of evolution prior to surgery, with isolated longitudinal lesion in the posterior horn of the medial meniscus, red-white area, without association with ligament lesion repaired with two sutures and two Fastener.

BERT models and hyperparameters

Library: Hugging Face 🤗

Spanish:

- **BETO:** *bert-base-spanish-wwm-uncased*
- **mBERT:** *bert-base-multilingual-cased*

English:

- **BioBERT:** *biobert-base-cased*
- **BERT:** *bert-base-cased*
- **mBERT:** *bert-base-multilingual-cased*

	Spanish		English		
	BETO	mBERT	BioBERT	BERT	mBERT
Batch size	8	16	8	16	18
Max Len	512	512	512	512	512
Learning Rate	2e-5	3e-5	2e-5	2e-5	3e-5
Epoch	3	5	4	5	5

Results

Language	System	Precision (%)	Recall (%)	F-score (%)
Spanish	BETO	54.61	69.59	61.20
	mBERT	58.76	50.23	54.16
English	BioBERT	66.25	60.60	63.30
	BERT	54.53	59.68	56.99
	mBERT	64.16	49.08	55.61

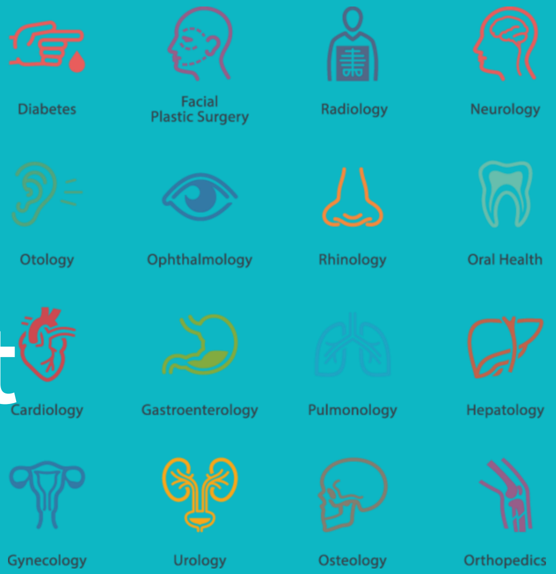
Other results using the training dataset:
https://github.com/plubeda/mie_2021/blob/main/Additional_results.md

Conclusion

- ❖ Moderate results
  - ➢ Complexity of the multi-label classification task
  - ➢ Cases may belong to several clinical disciplines, boundary decisions difficult
- ❖ Spanish dataset:
  - ➢ BETO performance better than mBERT
- ❖ English dataset:
  - ➢ BioBERT outperforms BERT
  - ➢ BioBERT outperforms any Spanish model (despite possible information loss)
  - ➢ Probable improvement if training models with biomedical data in Spanish (asBioBERT)
- ❖ To be done: use real clinical data for validation (less standardised language, less intricate clinical cases)

		F-score (%)
ES	BETO	61.20
	mBERT	54.16
EN	BioBERT	63.30
	BERT	56.99
	mBERT	55.61

Transfer learning for classifying Spanish and English text by clinical specialties



Alexandra POMARES-QUIMBAYA* Javeriana University, Bogotá, Colombia
Pilar LÓPEZ-ÚBEDA Jaén University, Spain
Stefan SCHULZ Medical University of Graz, Austria
*contact: pomares@javeriana.edu.co

31st Medical Informatics Europe Conference - May 31, 2021