Medical mT5: An Open-Source Multilingual Text-to-Text LLM for the Medical Domain

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata and Andrea Zaninello

iker.garciaf@ehu.eus

rodrigo.agerri@ehu.eus

https://hf.co/HiTZ/Medical-mT5-large https://hf.co/HiTZ/Medical-mT5-xl

Current Available Models for the Medical Domain

		# Param	Text2Text	Multilingual
XLM-RoBERTa	Conneau et al 2019	250M-12B	No	Yes
mDeBERTa-v3	He et al 2020	86M	No	Yes
BioBert	Lee et al 2019	110M	No	No
PubmedBERT	Gu et al 2020	110M	No	No
SciFive	Phan et al. 2021	220M-770M	Yes	No
BSC-BIO	Carrino et al 2022	125M	No	No
BioLinkBERT	Yasunaga et al 2022	110M-340M	No	No
BioT5X	Phan et al. 2022	110M-340M	Yes	No
BioGPT	Luo et al. 2022	347M	Yes	No
BioMedLM	Venigalla et al. 2022	2.7B	Yes	No
Med-PaLM	Singhal et al. 2022	540B	Yes	No
EriBERTa	To be published		No	Yes
Our Medical mT5		738M-3B	Yes	Yes

General domain

Medical Domain Corpus

Language	Sentences	Words
English	21.226.236	1.095.799.810
Spanish	35.312.809	960.055.764
French	7.192.779	670.972.717
Italian	3.504.555	143.164.133
Total	70.740.934	3.013.156.557

x2 (oversampling)

Medical Domain Corpus

Language	Sentences	Words
English	21.226.236	1.095.799.810
Spanish	35.312.809	960.055.764
French	7.192.779	670.972.717
Italian	3.504.555	143.164.133
Total	70.740.934	3.013.156.557
BioBERT	-	17.000.000.000
SciFive	-	78.643.200.000
mT5	-	1 trillion

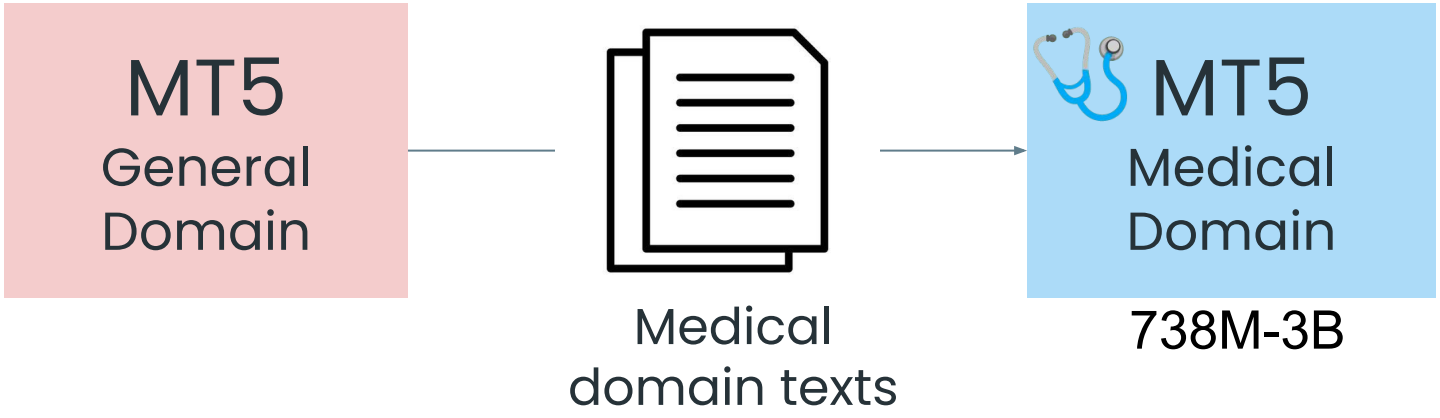
Medical Domain Corpus

Language	Sentences	Words
English	21.226.236	1.095.799.810
Spanish	35.312.809	960.055.764
French	7.192.779	670.972.717
Italian	3.504.555	143.164.133
Total	70.740.934	3.013.156.557
BioBERT	-	17.000.000.000
SciFive	-	78.643.200.000
mT5	-	1 trillion

Enough for fine-tuning a pre-trained mT5. Not enough to train an MT5 model from scratch

Model Training

We finetune mT5 checkpoints (model pre-trained with 1 trillion tokens from mC4)



Evaluation

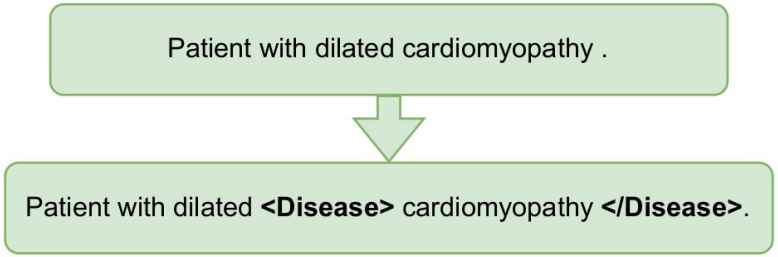


Figure 1: Text-to-Text representation of the Sequence Labeling task. Given an input sentence, the model must generate the same sentence annotated with html-style tags.

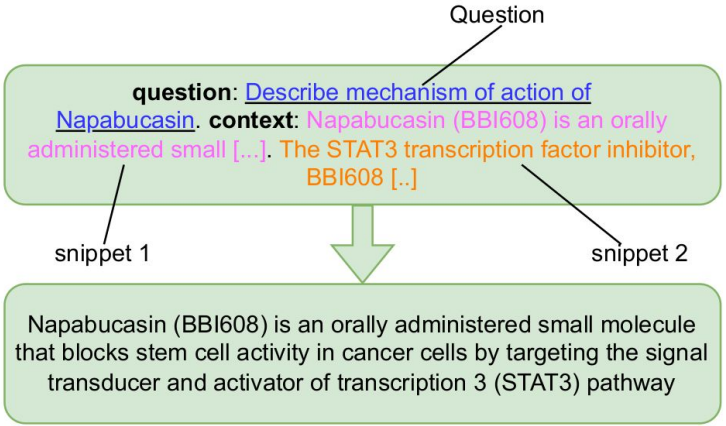


Figure 2: Text-to-Text representation of the BioASQ task. Given a question and a set of relevant snippets, the model must generate an answer.

Results - Sequence Labelling multitask

Lang	Dataset	Single Task FlanT5$_{XL}$	Single Task MedMT5$_{large}$	Single Task MedMT5$_{XL}$	MultiTask FlanT5$_{XL}$	MultiTask MedMT5$_{large}$	MultiTask MedMT5$_{XL}$
EN	NCBI-Disease	**89.3**	89.1	87.2	87.6	87.6	86.9
EN	BC5CDR Disease	**85.8**	84.4	82.4	85.1	83.4	83.0
EN	BC5CDR Chemical	**92.9**	92.8	91.3	92.7	92.5	91.6
EN	DIANN	74.2	74.8	77.6	**80.0**	75.4	75.3
ES	DIANN	70.9	74.9	74.8	**77.1**	72.6	73.6
EN	E3C	**63.1**	59.4	57.9	62.1	60.9	62.0
ES	E3C	67.1	72.2	69.5	66.5	**74.9**	73.3
FR	E3C	64.3	65.2	**65.8**	62.9	65.4	65.1
IT	E3C	65.1	**67.5**	65.9	60.7	66.9	65.1
ES	PharmaCoNER	89.1	**90.8**	90.1	89.9	90.3	89.5
EN	Neoplasm	73.4	**73.9**	73.2	73.1	72.3	72.9
EN	Glaucoma	**78.0**	76.2	76.4	76.4	76.8	77.5
EN	Mixed	**74.5**	72.2	72.0	71.5	70.9	73.0
ES	Neoplasm	**73.9**	72.1	71.8	73.5	73.5	73.7
ES	Glaucoma	75.2	77.1	75.5	77.1	77.7	**79.3**
ES	Mixed	71.6	72.4	71.4	70.0	71.8	**72.8**
FR	Neoplasm	73.7	72.9	71.2	**74.0**	72.9	73.6
FR	Glaucoma	77.2	**79.5**	75.8	76.6	77.0	79.4
FR	Mixed	**74.3**	73.3	69.7	71.8	71.2	73.0
IT	Neoplasm	72.0	71.2	73.1	71.9	**74.6**	74.0
IT	Glaucoma	75.9	75.7	78.7	77.6	78.5	**78.9**
IT	Mixed	70.0	70.6	71.9	69.9	72.5	**73.3**
	AVERAGE	75.1	75.4	74.7	75.2	76.2	**76.7**
	AVERAGE ES, FR, IT	72.9	74.0	73.2	73.1	74.8	**75.3**

Table 8: Multi-task supervised F1 scores for Sequence Labelling.

Results - Sequence Labelling Zero-shot

Lang	Dataset	$mT5_{XL}$	SciFive	$FlanT5_{XL}$	$mDeBERTa_{V3\ base}$	$MedMT5_{large}$	$MedMT5_{XL}$
ES	Neoplasm	71.4	69.8	67.9	65.1	**72.4**	71.7
ES	Glaucoma	**74.1**	71.5	70.6	68.3	72.4	73.2
ES	Mixed	**69.4**	67.0	66.7	60.9	68.1	68.8
FR	Neoplasm	71.6	68.6	69.9	60.5	72.4	**72.8**
FR	Glaucoma	75.8	74.5	71.0	68.7	72.3	**76.7**
FR	Mixed	**73.0**	68.5	68.2	59.3	70.4	72.4
IT	Neoplasm	70.6	63.1	67.3	62.4	72.9	**73.2**
IT	Glaucoma	76.7	71.6	72.0	70.2	75.4	**79.0**
IT	Mixed	69.9	62.5	66.9	62.1	71.7	**71.9**
AVERAGE		72.5	68.6	69.0	64.2	72.0	**73.3**

Table 9: Zero-shot F1 scores for Argument Mining. Models have been trained in English and evaluated in Spanish, French and Italian.

Results - Expert Manual Evaluation

Ideal Answer	Which type of urinary incontinence is diagnosed with the Q tip test? Stress urinary incontinence is diagnosed with the Q tip test. The test evaluates urethral mobility.	Ranked
mT5-large	Q-tip test is used for urodynamic assessment of patients with severe urinary incontinence.	4º
FlanT5-large	The Q tip test is used for assessing urethral mobility.	3º
SciFive	Q tip test is used for evaluation of stress urinary incontinence.	2º
MedMT5-large	Q-tip test is used for evaluation of stress urinary incontinence.	1º
Ideal Answer	Which are the main manifestations of Ohdo syndrome? Severe ID, absent or deficient language, skeletal manifestations including bilateral patella dislocations.	Ranked
mT5-large	Skeletal manifestations in Ohdo syndrome are a case with bilateral patella dislocations where surgical intervention has been indicated.	4º
FlanT5-large	The main manifestations of Ohdo syndrome are: 1) severe ID, 2) absent or deficient language and 3) milder, clinical manifestation in heterozygotes.	2º
SciFive	Ohdo syndrome is characterized by severe ID, absent or deficient language and, milder, clinical manifestation in heterozygotes.	1º
MedMT5-large	The main manifestations of Ohdo syndrome are: 1) absent or deficient language and 2) mildder clinical manifestation in heterozygotes.	3º

Table 8: Examples of answers generated by each model for two different BioASQ questions together with the rank assigned by medics.

Summary and remaining work

- Two MedMT5 models trained for English, Spanish, French and Italian with 770M and 3B parameters
  - https://huggingface.co/HiTZ/Medical-mT5-xl
  - https://huggingface.co/HiTZ/Medical-mT5-large
- New datasets for Argument Mining and QA for French, Italian and Spanish
  - https://huggingface.co/datasets/HiTZ/multilingual-abstrct
  - https://huggingface.co/datasets/HiTZ/Multilingual-BioASQ-6B
- Largest publicly available medical multilingual corpus for Spanish, French, and Italian languages (3B tokens)
  - https://huggingface.co/datasets/HiTZ/Multilingual-Medical-Corpus
- Large models: LlaMA (7B, 13B, and 35B) should help to improve results
- Text2Text Evaluation: remains ad-hoc and unclear

Antidote Project: Huggingface data collection

https://hf.co/collections/HiTZ/antidote-project-6601973d7d7b55302c1e606d

Or find it at:

- https://hf.co/HiTZ

Medical mT5: An Open-Source Multilingual Text-to-Text LLM for the Medical Domain

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata and Andrea Zaninello

iker.garciaf@ehu.eus

rodrigo.agerri@ehu.eus

https://hf.co/HiTZ/Medical-mT5-large https://hf.co/HiTZ/Medical-mT5-xl