

Reading and interpreting human language to extract knowledge: a challenge of Artificial Intelligence at the service of Emergency Medicine



*Bernardo Magnini
Bruno Kessler Foundation, Trento, Italy
magnini@fbk.eu*

November 3, 2022
BOLOGNA, IT



Funded by
the European Union

eCREAM project
N. 1010557726



Outline

- Natural Language Processing (NLP) and Artificial Intelligence
 - NLP: big data and machine learning, current challenges
- Core technologies in NLP: language modeling
 - Language models, neural language models
- Transformers: a big step forward
 - Continuous pre-training, fine-tuning
- Scientific challenges for eCream
 - Multilingual pre-trained models for the medical domain



Natural Language Processing and Artificial Intelligence

- We are in the era of “big data” and machine learning
 - Natural Language Processing (NLP) is not an exception!
- NLP requires massive amounts of data for training NLP models
 - Machine translation, sentiment analysis, chatbots, **information extraction**, summarization, etc.

Why NLP for the Medical Domain?

- A significant amount of information is still in textual format
 - Need to be extracted and stored
 - High language variability: different terminologies, different languages
 - Non grammatical language, acronyms, abbreviations, typing errors
- Few examples
 - Classification of clinical reports
 - Coding ICD-10 pathologies
 - Extract relevant information

<i>Present complaint</i>
Male, 79y. Presents after a loss of consciousness during the night in standing position post-micturition. Recalls a sudden onset of lightheadedness just before loosing consciousness. Reports a mild head injury in the fall. Upon awakening, patient called 999 on his own. No PTA. Does not report recent similar episodes.
<i>Social history</i>
Pt lives alone, caregiver a few hours a day.
<i>Past medical history</i>
Hypertension, benign prostatic hyperplasia, mild chronic kidney disease, CHD (PTCA).
<i>Drug history</i>
tamsulosin, enalapril, ASA 100 mg, statin, bisoprolol, lormetazepam. NKDA

Which data do we need?

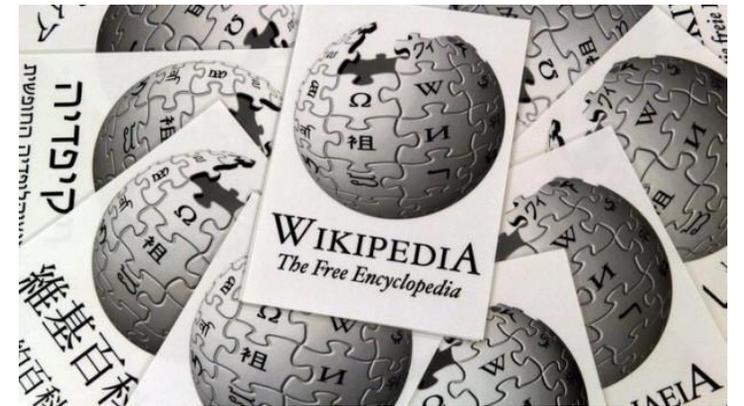
- Annotated data

- Require human supervision (domain experts), learning is expensive
- 10K+ annotations for a downstream task (training, dev, testing): classify clinical reports, extract diagnosis
- Issues: category unbalance, poor agreement among annotators, noisy data

- Non annotated data

- Do not require human supervision, learning is cheaper
- Used for language modeling, clustering (topic modeling)
- 100M words for language modeling
- Issues: noisy data, limited availability (e.g., data protection regulations)

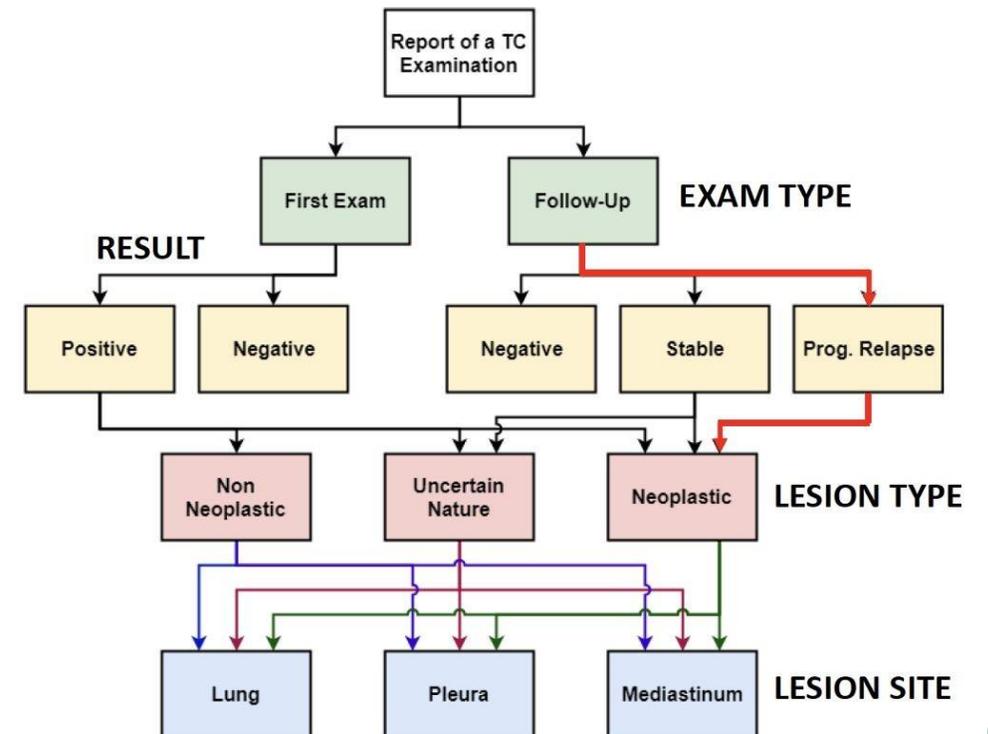
A 25-year-old man with a history of Klippel-Trenaunay syndrome presented to the hospital with mucopurulent bloody stool and epigastric persistent colic pain for 2 wk.



Current challenges in NLP

- Find a tradeoff between supervision and performance
 - Goal: reduce as much as possible the need of human annotated data
- Example: classification of radiological reports

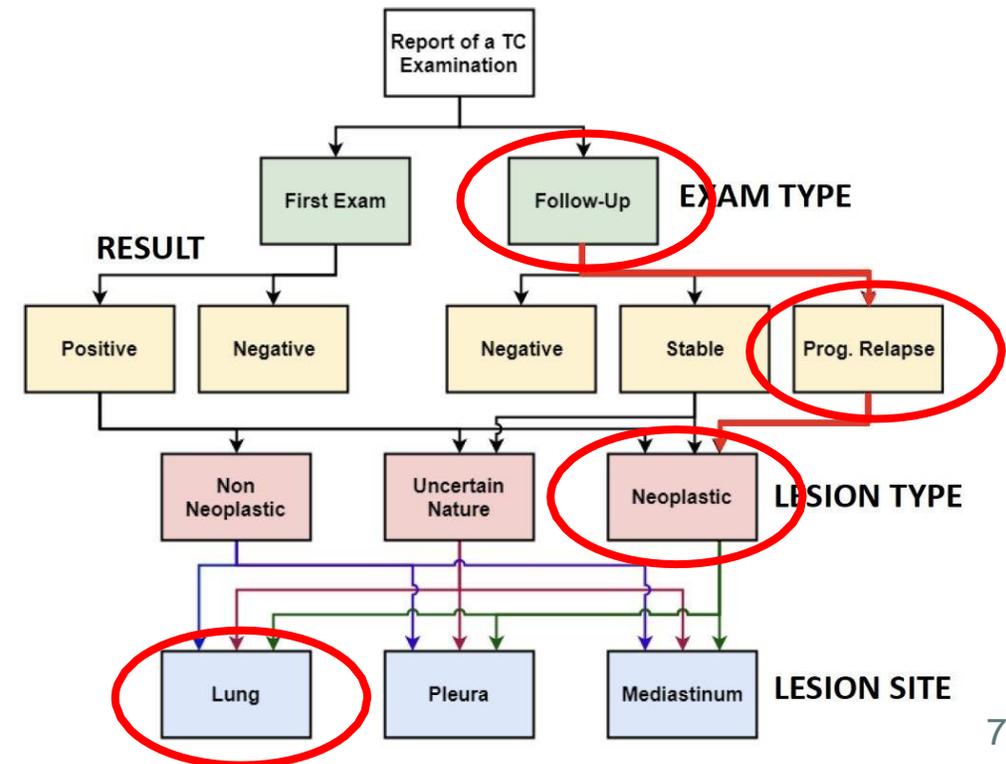
TC TORACE CON/SENZA MDC
Esame eseguito con somministrazione di 90 ml di Iomeron 350, confrontato con precedente del 7/6/2017. Al controllo attuale nel segmento posteriore del lobo superiore destro è riconoscibile **lesione nodulare di 15 x 15 mm con margini irregolari e spiculati**, localizzata in sede peri-ilare, a stretto contatto con sottili diramazioni vascolari e una diramazione bronchiale subsegmentaria che presenta pareti ispessite; a giudizio clinico utile broncoscopia. **Incremento del versamento pericardico.**



A Big Challenges for NLP

- Find a tradeoff between supervision and performance
 - Goal: reduce as much as possible the need of human annotated data
- Example: classification of radiological reports

TC TORACE CON/SENZA MDC
Esame eseguito con somministrazione di 90 ml di Iomeron 350, confrontato con precedente del 7/6/2017. Al controllo attuale nel segmento posteriore del lobo superiore destro è riconoscibile **lesione nodulare di 15 x 15 mm con margini irregolari e spiculati**, localizzata in sede peri-ilare, a stretto contatto con sottili diramazioni vascolari e una diramazione bronchiale subsegmentaria che presenta pareti ispessite; a giudizio clinico utile broncoscopia. **Incremento del versamento pericardico.**



Language Models

- A collection of probabilities derived from a collection of texts
 - The probability of a sequence of words (a sentence)
 - The probability of the next word, after a sequence of words

$$P(w_i | w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

the probability that "to" occurs after "want"

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Why are Language Models Useful?

- There is a lot of implicit knowledge about language and world!
- This knowledge is crucial for interpreting natural languages
- No need of human supervision!

$$P(\text{english} | \text{want}) = .0011$$

cultural knowledge

$$P(\text{chinese} | \text{want}) = .0065$$

$$P(\text{food} | \text{chinese}) = .052$$

word association,
terminology

$$P(\text{to} | \text{want}) = .66$$

syntactic knowledge

$$P(\text{eat} | \text{to}) = .28$$

$$P(\text{food} | \text{to}) = 0$$

$$P(\text{want} | \text{spend}) = 0$$

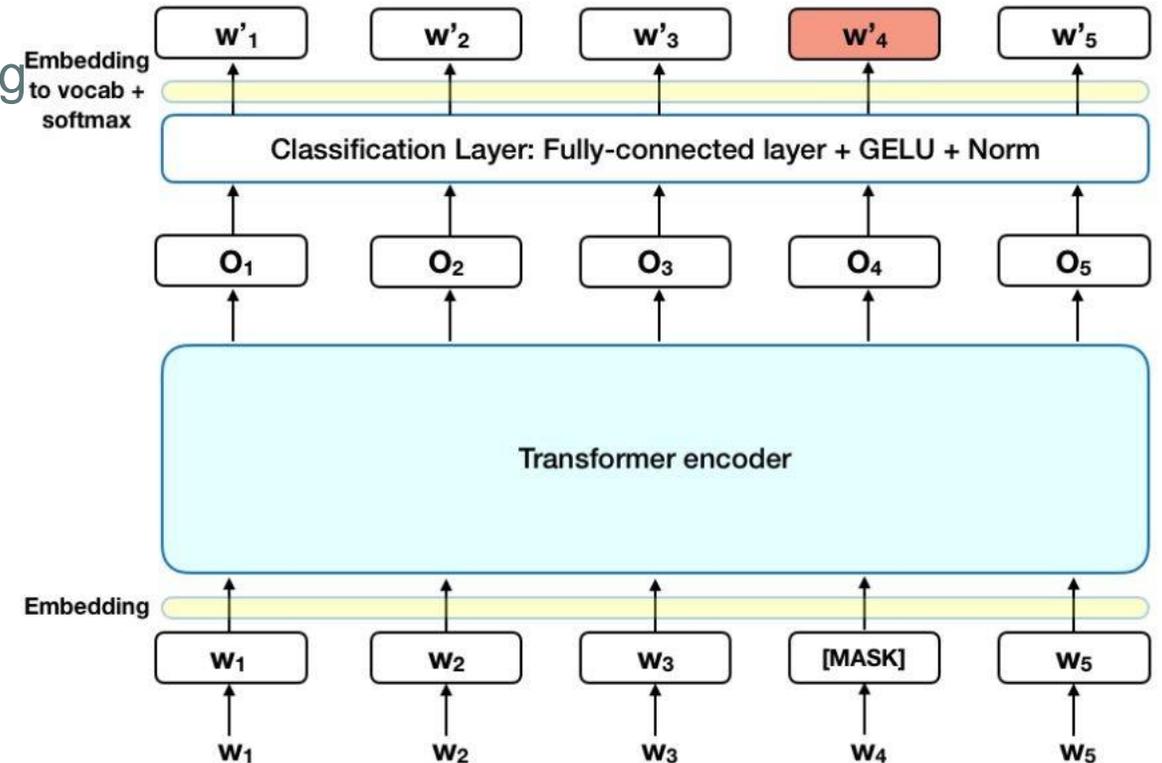
$$P(i | \langle s \rangle) = .25$$

knowledge about the
application

A Step Forward: Neural Language Models

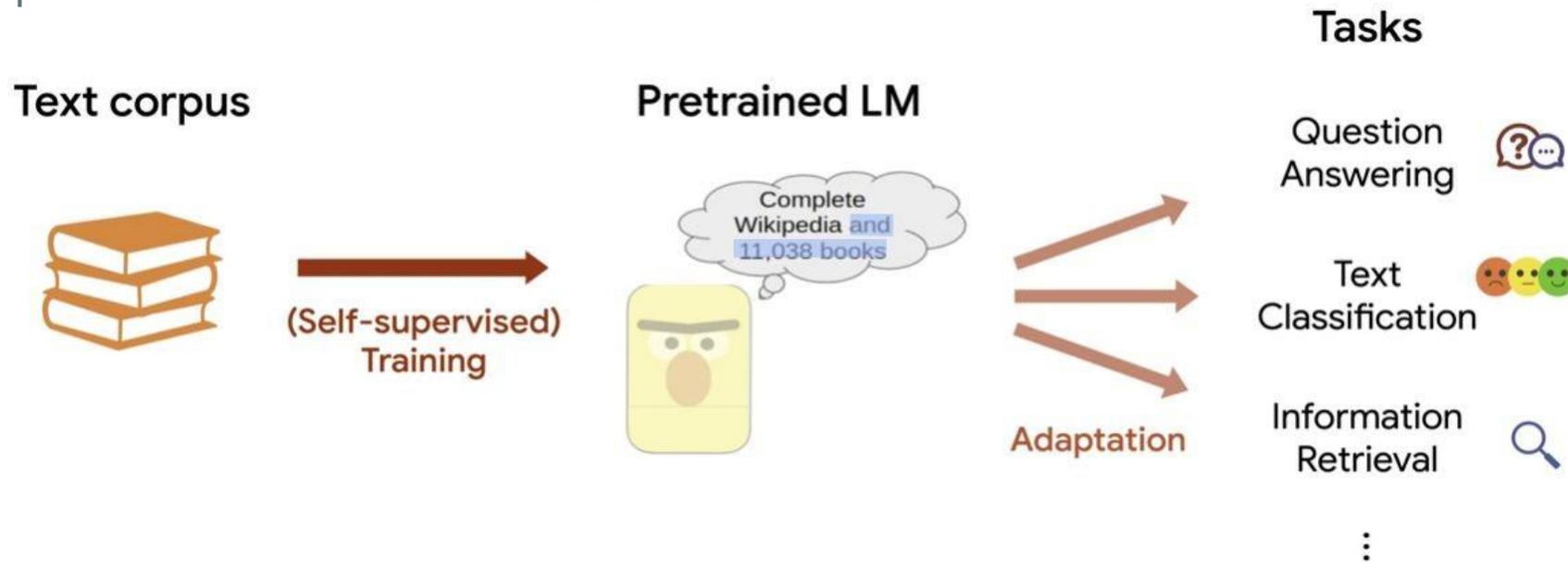
- Transformers

- **Predict** word probabilities, rather than counting them: better results (Baroni, et. al. 2014)
- Consider both **left and right context**
- Trained to guess the next word in a sentence, or a **masked word in a sentence**
- **Self-supervision**: do not require human annotations
- BERT (Google), GPT-3 (OpenAI), T5 (Google), and many others
- **Trained over 5BN+** words on English (e.g., Wikipedia)



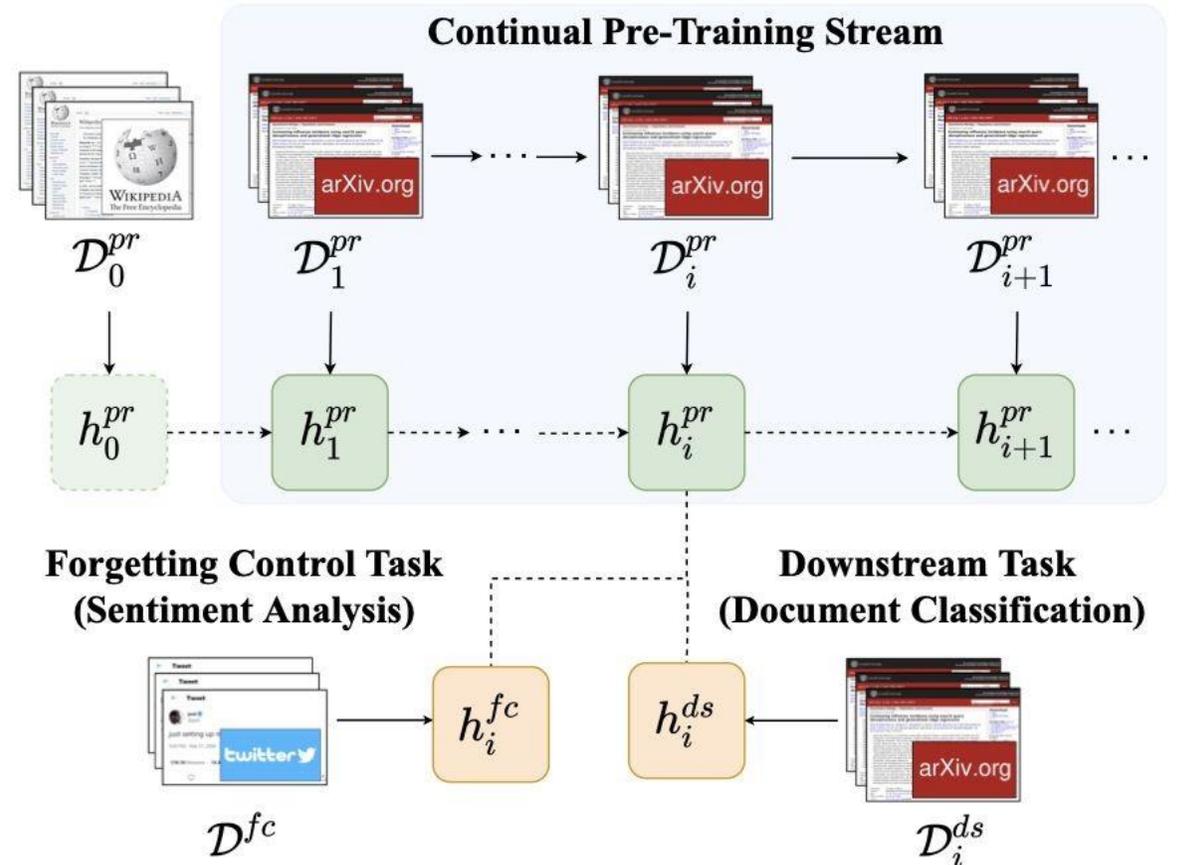
Using Neural Language Models

- Specialize a generic language model on a specific task
 - **Fine-tuning** on a specific task (downstream task)
 - Different objective function (e.g. document classification)
 - Require annotated data for the task



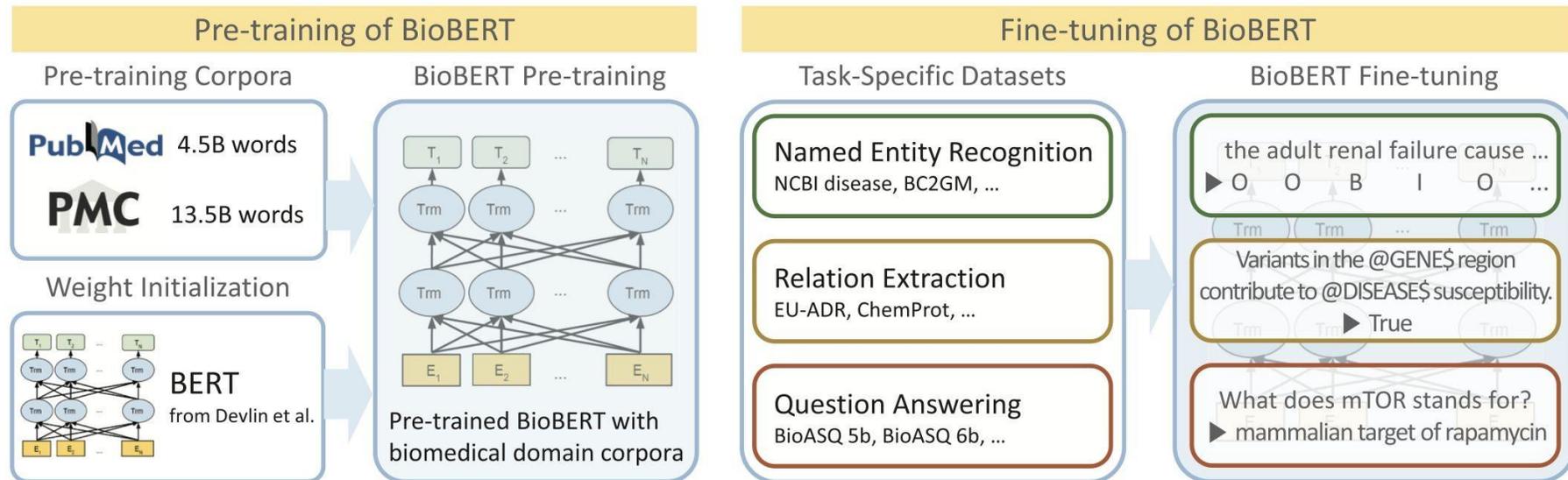
Continuous Pre-training of a Generic Language Model

- Continuous pre-training
 - Same objective function of the pre-training step (masked word)
 - Use as much as possible documents of a specific domain/genre (e.g., conversations, tourisms, medicine)



Pre-training on the biomedical Domain

- A “native” language model for the biomedical domain
 - Pretraining on biomedical corpora on the same BERT objective function (masked word)
 - BioBERT (Lee et al. 2019): state-of-the-art results on specific biomedical tasks



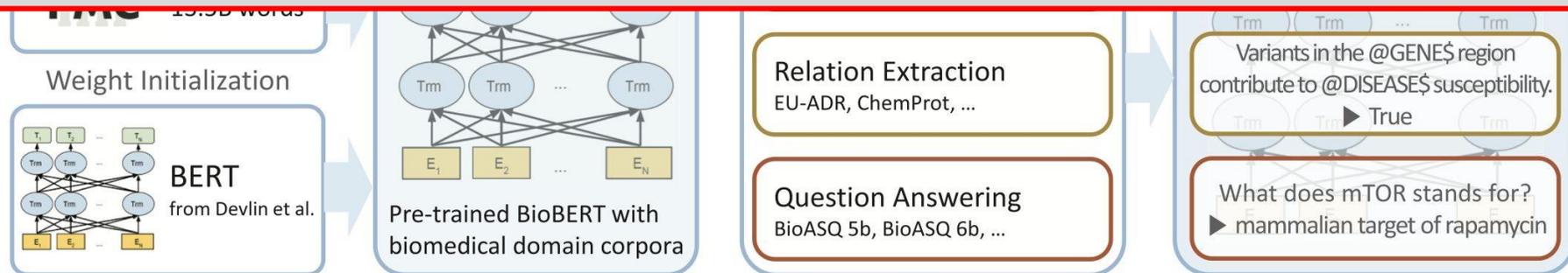
Pre-training on the biomedical Domain

- A “native” language model for the biomedical domain
 - Pretraining on biomedical corpora on the same BERT objective function (masked word)

This is for English !

Can we do something similar for other European languages?

A grand challenge for eCREAM (and not only)



Ongoing Work for eCREAM: mT5

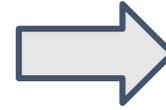
- A biomedical language model for Italian based on T5
 - 143M word (done)
 - Building mT5 for Italian (ongoing)
 - Test the model on eCREAM tasks
 - Extend to other eCREAM languages: a multilingual medical T5

67.048.457	commoncrawl_medico.txt
30.592.830	foglietto.txt
13.294.320	wikipedia_medicina.txt
11.630.234	e3c.txt
6.305.658	farmaci.txt
5.827.020	tesi.txt
2.287.728	pubmed.txt
1.346.562	integratori.txt
1.339.226	nurse24.txt
1.223.656	iss.txt
975.585	appunti.txt
904.215	humanitas.txt
489.977	mypst.txt
157.147	patologie.txt
26.553	simulazioni.txt
20.469	casiclinici.txt
143.469.637	

Fine Tuning mT5 on eCREAM Tasks

<i>Present complaint</i>
Male, 79y. Presents after a loss of consciousness during the night in standing position post-micturition. Recalls a sudden onset of light-headedness just before losing consciousness. Reports a mild head injury in the fall. Upon awakening, patient called 999 on his own. No PTA. Does not report recent similar episodes.
<i>Social history</i>
Pt lives alone, caregiver a few hours a day.
<i>Past medical history</i>
Hypertension, benign prostatic hyperplasia, mild chronic kidney disease, CHD (PTCA).
<i>Drug history</i>
tamsulosin, enalapril, ASA 100 mg, statin, bisoprolol, lorazepam. NKDA

<i>On examination</i>
Airway - patent Breathing - Sat 98%, RR 16, chest clear, no shortness of breath Circulation – BP 130/70, HR 82, WWP. Normal S1 and S2 heart sounds. No oedema. Disability – GCS 15, pupil, equal, round, reactive to light and accommodation, cranial nerve 2-12 intact, 5/5 strength in all extremities bilaterally Expose – T 35.8. Abdomen soft non tender. No other sign of trauma, no c-spine pain, no pelvic pain. EKG: sinus rhythm, 80 bpm, old inferior q wave, non specific ST-T changes. Cardiac POCUS: EF approximately 30%.
Pt alert, asymptomatic 12h negative cardiac telemetry Pt reports recent introduction of tamsulosin
<i>Conclusion</i>
Vasovagal syncope



<u>DIMENSION TO EXPLORE</u>	<u>VARIABLES TO RETRIEVE</u>	<u>VALUES TO BE EXTRACTED BY NLP</u>
Patient frailty	Sex and Age	Male, 79 years
	Functional status	Lives alone, caregiver a few hours a day
	Comorbidity	Hypertension, Benign Prostatic Hyperplasia, Chronic Kidney Disease, Coronary Artery Disease (PTCA)
	Drug history	tamsulosin, enalapril, ASA, statin, bisoprolol, lorazepam, NKDA
Syncope's features	Prodromal symptoms	Light-headedness
	Trigger	Post-micturition
	Conditions upon awakening	Aware (deduced), no PTA
	Palpitations	No (deduced)
Consequences	Brain injury	Mild, without bleedings
	Fractures	No
Exams	EKG	Sinus rhythm, 80 bpm, old inferior q wave, non-specific ST-T changes
	Lab	No
	Echocardiography	EF approximately 30%
	Monitor	Yes

Entity extraction, identifying relevant information in text (such as the presence of signs and symptoms, suspected and confirmed diagnosis, anamnesis).

Entity linking, linking such relevant information to corresponding coding repositories (e.g., ICD-9-CM, UMLS).

TAKE HOME MESSAGE

- **Neural language models (transformers)** are the current state of the art in Natural Language Processing
- Challenge for eCREAM: **a native language model for the biomedical domain** for European languages
- We need documents for the biomedical domain!
 - Scientific papers, theses, clinical documents, etc.
- We need annotated documents for downstream eCREAM tasks



THANK YOU
FOR YOUR
ATTENTION

Transformers

- **Efficiency:** trained on several billions
- Take advantage of **attention mechanism**
- Consider both **left and right context**
- Contextualized representation of word meaning

