

PyPharma NLP Tutorial 2019: Learning by doing NLP

Diego Saldana, PHC Analytics

November 2019

Agenda

- PyPharma Conference 2019
- Natural Language Processing
- PyPharma NLP 2019
- Notebooks Format
- Examples
 - Masked Language Modelling
 - Detecting Adverse Events in the Literature
 - Extracting Disease Mentions
 - Extracting Gene-Disease Associations
 - Question Answering

- Aimed at **python** users in the pharmaceutical **industry** and **academia**, all aspects of the pharmaceutical **lifecycle** and data **modalities**.
- Attendance is estimated at around **100** people (still growing).
- **Attendants** and **speakers** from Roche, Novartis, GSK, AstraZeneca, UNIBAS, UNIL, IBM, ETHZ, UZH, SIB, and various other companies and universities.

- Two days, **single** track, **invite** only, **free** to attend.
- **Hosts:** Roche, Novartis, and the University of Basel.
- Day 1: November 21st at the **University of Basel**, Day 2: November 22nd at the **Roche** Viaduktstrasse amphitheater.

Natural Language Processing (1/2)

- **Automate** the processing of data in **natural language** form.
- Look at **text** as a **database** that we want to be able to **query** in various ways.
- **Combines** elements of linguistics, computer science, statistics, artificial intelligence, etc.
- One natural way to do this is to use **machine learning** to perform **tasks** such as text classification, named entity extraction, relation extraction, etc.

Natural Language Processing (2/2)

- In recent years, **deep learning** based methods have shown great promise in terms of performance.
- But also the ability to **transfer knowledge** across datasets as well as across tasks.
- Examples include models such as
 - word2vec
 - GloVe
 - ELMO
 - BERT
 - etc

Natural Language Processing (3/2)

- The usual **transfer learning** procedure is
 - Step 1: train a **base model** to perform **generic** tasks on very **large** datasets (websites, books, wikipedia, social media, etc).
 - Step 2: **fine tune** the model to perform a **specific** task on a **smaller** dataset.
- **Extensions** to these methods to **biomedical** applications often follow (e.g. Pyysalo embeddings for word2vec, BioBERT for BERT, etc).

- A **tutorial** on Pharma and Biomedical NLP that will take place on Day 1 of PyPharma (November 21st at the University of Basel).
- The goal is for intermediate pharmaceutical and biomedical python users to **learn** how to do **state of the art** NLP by **doing** it.
- Since doing it is usually **hard**, we provide **tools** for them to make the process **easier**.
- We provide **notebooks** with examples of various **tasks** and **datasets**.

Notebooks: Format (1/2)

- They will be hosted in **Google Colab** as well as Azure Notebooks.
- **No need** to install anything in your computer, or buy a new one, to do **state of the art**, deep learning based NLP (yay!).
- Prior to using the notebooks, an **introduction** to deep learning based NLP will be done.

- The common **structure** will be:
 - **Downloading** the training datasets.
 - Exploring the **training data** and how it is seen by the model (**inputs** and **outputs**).
 - Training the model and **storing** the results in a **checkpoint**.
 - Re-loading the checkpoint and performing **predictions** on new data, as well as **exploring** the results **interactively**.

Examples: Masked Language Modelling

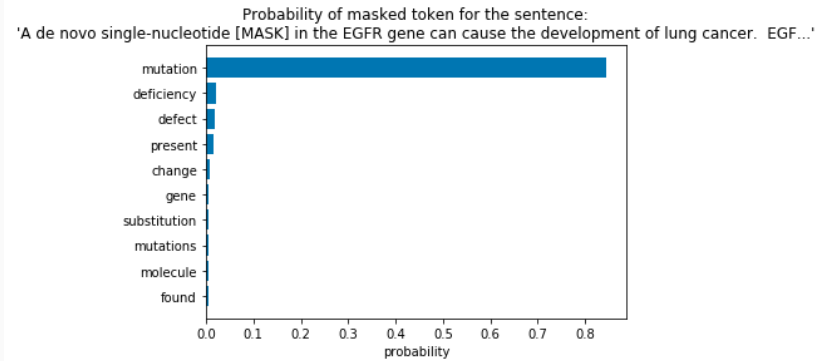


Figure 1:

Examples: Detecting Adverse Events

	Sentence Label
INTRODUCTION: Neuropsychiatric symptoms are a well-described side effect of systemic corticosteroid therapy and can range from mild to severe.	Neg
CASE PRESENTATION: We describe a case of substance-induced psychosis following epidural injection of 10 mg dexamethasone.	AE
Three days after the procedure, the patient developed symptoms including anger, hostility, insomnia, paranoia, and delusions.	Neg
Symptoms resolved between 7 and 17 days.	Neg
DISCUSSION: In the past 50 years, there have been several case reports of severe neuropsychiatric effects following intraarticular or other interventional pain injections with various corticosteroids.	Neg
More recent reviews have identified possible risk factors, including corticosteroid dose, patient age, sex, and history of neuropsychiatric disorder, among others, although these conclusions are not duplicated across all studies.	Neg
CONCLUSION: Recommendations for practice include patient and family education on possible adverse effects of corticosteroid administration, utilization of minimum effective doses for interventional procedures, and the consideration of close follow-up and multidisciplinary coordination, especially in high-risk patients.	Neg

Figure 2:

Examples: Extracting Disease Mentions

Sentence ID	Token	True Label	Predicted Label
13	None	O	O
13	of	O	O
13	the	O	O
13	patients	O	O
13	had	O	O
13	decompensated	O	O
13	liver	B	B
13	disease	I	I
13	.	O	O

Figure 3:

Examples: Extracting Gene-Disease Associations

Our data shows no association between @DISEASE\$ and the Leu432Val polymorphism of the @GENE\$ gene or the tetranucleotide repeats of the CYP19 gene.	1	1
Our study showed that gene polymorphisms of @GENE\$ and SULT1A1 induce an individual susceptibility to @DISEASE\$ among current smokers.	1	1
We conclude that the @GENE\$*3 allele appears to be a factor for susceptibility to @DISEASE\$ in Turkish women especially those with a BMI greater than 24 kg/m(2).	1	1
Our results suggested that the Val @GENE\$ allele increases the susceptibility to @DISEASE\$ in women exposed to waste incinerator or agricultural pollutants.	1	1
These results do not support a favoring role of @GENE\$*3 in @DISEASE\$ development in our population.	1	1
We found no evidence for an overall association between @GENE\$ genotype and @DISEASE\$ risk, nor was there any clear indication of gene-environment interaction.	1	1
These results suggest that HLA class I antigens and @GENE\$ A-308G are not associated with susceptibility or resistance to the development of TDI-induced @DISEASE\$.	0	1
These results suggest that the C1772T polymorphism in @GENE\$ is not involved in @DISEASE\$ or metastasis of colorectal carcinoma.	0	0
These results provide substantial evidence that genetic variation within or extremely close to @GENE\$ impacts both @DISEASE\$ risk and traits related to the severity of AD.	0	0
This study provides the first evidence that @GENE\$ may be a candidate susceptibility loci that affects the @DISEASE\$ of atherosclerosis in Japanese subjects.	0	0
Since only eight out @GENE\$ iron-overloaded HbH patients carry a @DISEASE\$ in the TFR2 or HFE gene in the heterozygote state and their iron loading status was comparable to the matched controls without such defects, it would appear that the accumulation of excess iron in HbH disease is more likely a result of increase dietary absorption secondary to ineffective erythropoiesis.	0	0
The H63D @DISEASE\$ of the @GENE\$ gene has a moderate but significant influence on sTfR concentration in the general population, the presence of one or two mutated alleles being associated with an average of 0.27 mg/L less sTfR than nonmutated homozygotes.	0	0

Figure 4:

Examples: Question Answering (1/2)

Importance: Osimertinib mesylate is used globally to treat EGFR-mutant non-small cell lung cancer (NSCLC) with tyrosine kinase inhibitor resistance mediated by the **EGFR T790M mutation**. Acquired resistance to osimertinib is a growing clinical challenge that is poorly understood. Objective: **To understand the molecular mechanisms of acquired resistance to osimertinib** and their clinical behavior. Design, Setting, and Participants: Patients with advanced NSCLC who received osimertinib for T790M-positive acquired resistance to prior EGFR tyrosine kinase inhibitor were identified from a multi-institutional cohort (n = **143**) and a confirmatory trial cohort (NCT01802632) (n = **110**). Next-generation sequencing of tumor biopsies after osimertinib resistance was performed. Genotyping of plasma cell-free DNA was studied as an orthogonal approach, including serial plasma samples when available. The study and analysis were finalized on **November 9, 2017**. Main Outcomes and Measures: **Mechanisms of resistance and their association with time to treatment discontinuation on osimertinib**. Results: Of the 143 patients evaluated, 41 (**28 [68%]** women) had tumor next-generation sequencing after acquired resistance to osimertinib. Among 13 patients (32%) with maintained T790M at the time of resistance, EGFR C797S was seen in 9 patients (22%). Among 28 individuals (68%) with loss of T790M, a range of competing resistance mechanisms was detected, including novel mechanisms such as **acquired KRAS mutations and targetable gene fusions**. Time to treatment discontinuation was shorter in patients with T790M loss (6.1 vs 15.2 months), suggesting emergence of pre-existing resistant clones; this finding was confirmed in a validation cohort of 110 patients with plasma cell-free DNA genotyping performed after osimertinib resistance. In studies of serial plasma levels of mutant EGFR, loss of T790M at resistance was associated with a smaller decrease in levels of the EGFR driver mutation after 1 to 3 weeks of therapy (100% vs 83% decrease; P = .01). Conclusions and Relevance: Acquired resistance to osimertinib mediated by loss of the T790M mutation is associated with early resistance and a range of competing resistance mechanisms. These data provide clinical evidence of the heterogeneity of resistance in advanced NSCLC and a need for clinical trial strategies that can overcome multiple concomitant resistance mechanisms or strategies for preventing such resistance.

Figure 5:

Examples: Question Answering (2/2)

QUESTION:

What mediates TKI resistance? EGFR T790M mutation.

ANSWER: tyrosine kinase inhibitor resistance mediated by the EGFR T790M mutation

What are the main outcomes of this study?

ANSWER: Mechanisms of resistance and their association with time to treatment discontinuation on osimertinib

How many patients participated in this study?

ANSWER: 143

What are some competing resistance mechanisms to T790M?

ANSWER: acquired KRAS mutations and targetable gene fusions

What is the objective of the study?

ANSWER: To understand the molecular mechanisms of acquired resistance to osimertinib

When was the study finalized?

ANSWER: November 9, 2017

How many of the patients in this study were women?

ANSWER: 28 [68%]

Figure 6: