PyPharma NLP Workshop 2019

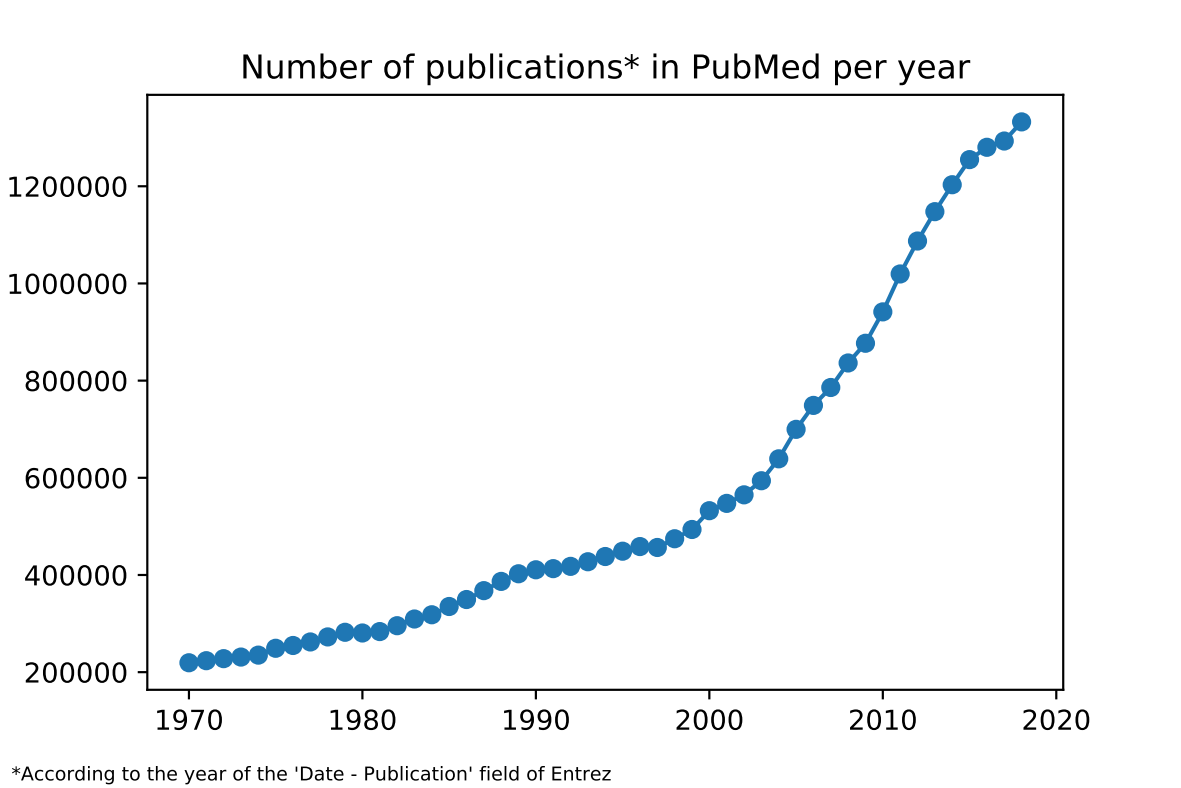
Diego Saldana

November 2019

Introduction to Biomedical NLP

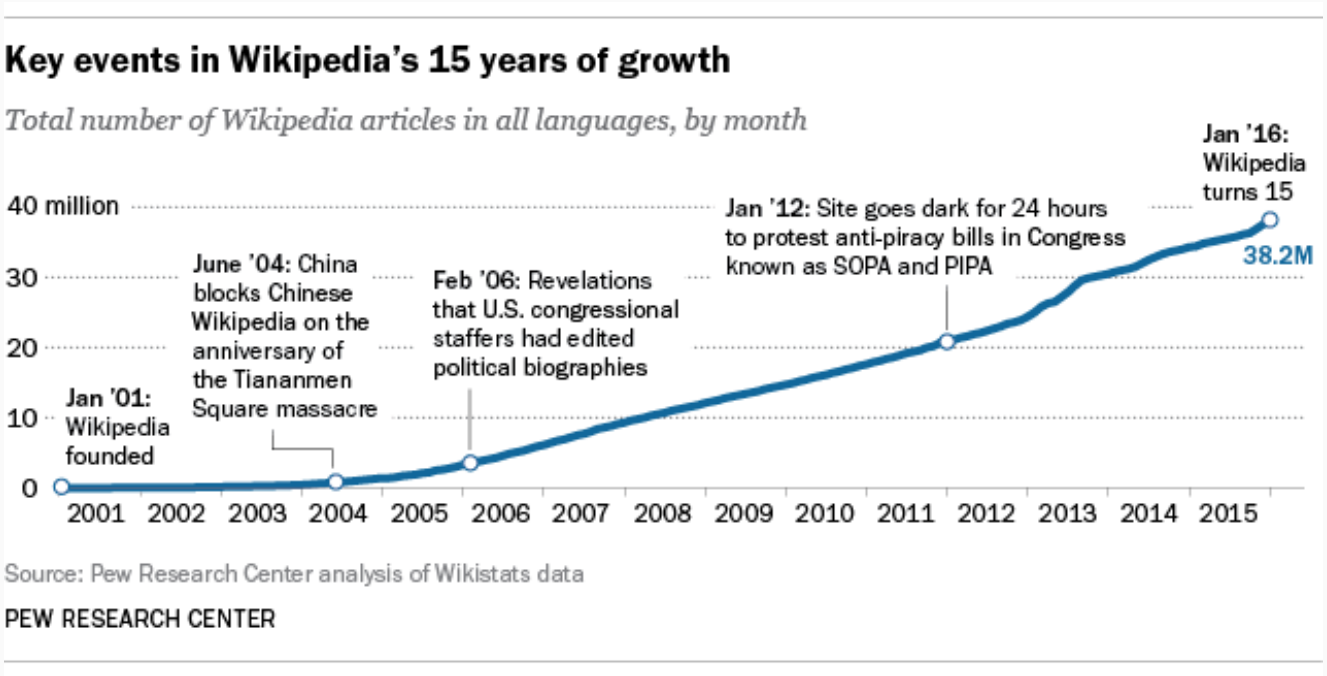
Why Biomedical NLP? (1/3)

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.



Why Biomedical NLP? (2/3)

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.



Why Biomedical NLP? (3/3)

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.
- This information is potentially very useful but cannot readily be used programmatically and stored in databases, searched, or analyzed.
- As a result this valuable information is "locked into a vault" until a human reads it, structures it and puts it into some database.
- And even when that happens, the scope in which the data can be used is usually limited and chosen by the extractors.
- How can machines help?

Humans vs. Machines (1/2)

- Machines and humans have different strengths and weaknesses when processing text.

Table 5. IAA scores between the annotators over the *ADE-seed-set1* corpus containing 50 documents. Enumerations related to dosages are zeroes since no dosage information was annotated during this round.

Annotators	**Entity (*exact match*)**			**Entity (*partial match*)**		
	Drug	**Adverse effect**	**Dosage**	**Drug**	**Adverse effect**	**Dosage**
1 and 2	0.76	0.66	0.00	0.82	0.86	0.00
1 and 3	0.28	0.43	0.00	0.38	0.55	0.00
2 and 3	0.29	0.40	0.00	0.38	0.51	0.00

Annotators	Relation (*exact entity match with exact relation*)		Relation (*partial entity match with exact relation*)	
	Drug-adverse effect	Drug-dosage	Drug-adverse effect	Drug-dosage
1 and 2	0.64	0.00	0.79	0.00
1 and 3	0.14	0.00	0.37	0.00
2 and 3	0.10	0.00	0.37	0.00

Humans vs. Machines (2/2)

- Machines and humans have different strengths and weaknesses when processing text.
- Machines in particular are capable of processing vast amounts of text in a very short period of time in a very consistent way and performing simple tasks.
- Humans are take much more time to process text and are less consistent, however they are capable of much more complex reasoning and understanding.

Humans vs. Machines (3/2)

What are some examples of tasks can computers perform well in 2019?

- Categorizing documents (e.g. automatically assigning MeSH headings to PubMed abstracts)
- Extracting entities from text (e.g. extracting Drugs, Diseases from PubMed abstracts)
- Extracting relations from text (e.g. extracting Adverse Events from PubMed abstracts)
- Answering simple questions based on a small amount of context (e.g. "Which drug should be used as an antidote in benzodiazepine overdose?")

Some Common Tasks

Language Modelling: A language model assigns probabilities to sequences of tokens, where tokens t can be words, characters, sub-words, etc:

$$P\Big(t_1, t_2, t_3, ..., t_N\Big).$$

One common way to do this is to decompose this as the probability of the next token in the sequence t_i given the probability of the sequence up to the previous token and some parameters Θ for our model:

$$P\Big(t_i | t_1, t_2, t_3, ..., t_{i-1}, \Theta\Big).$$

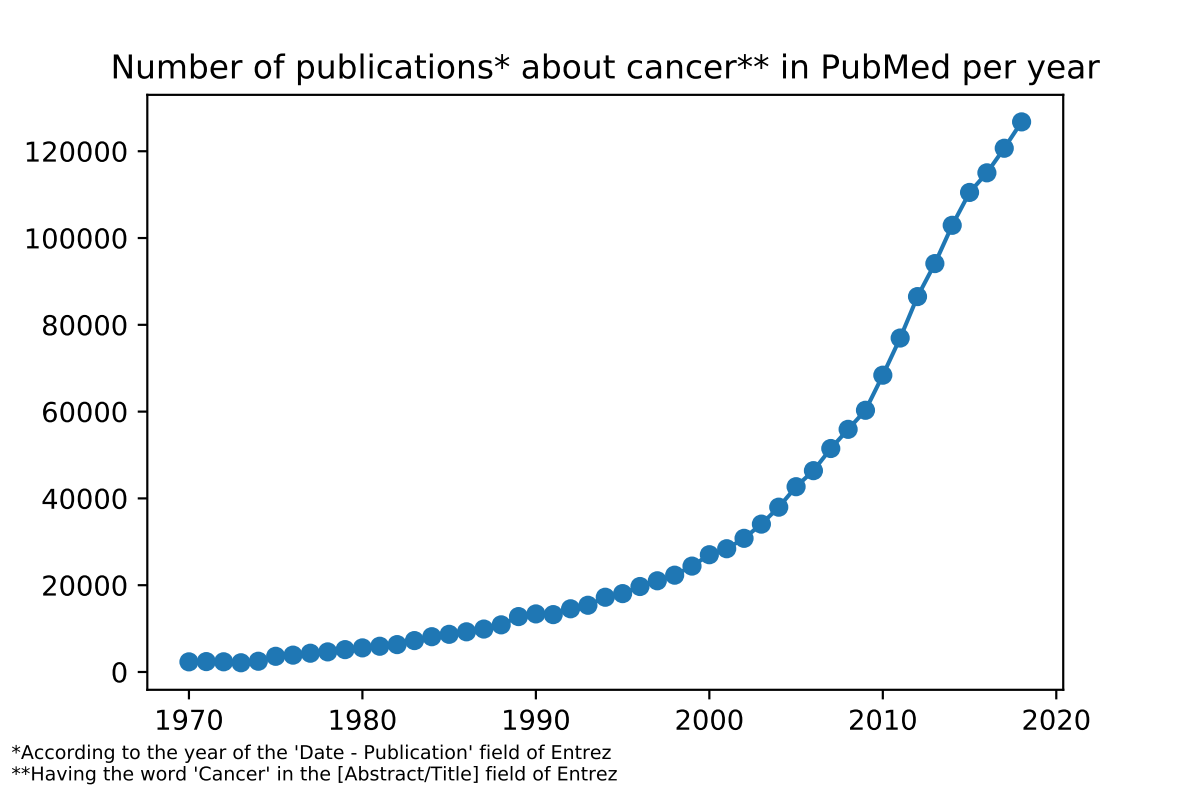
References

WIP.

Backup

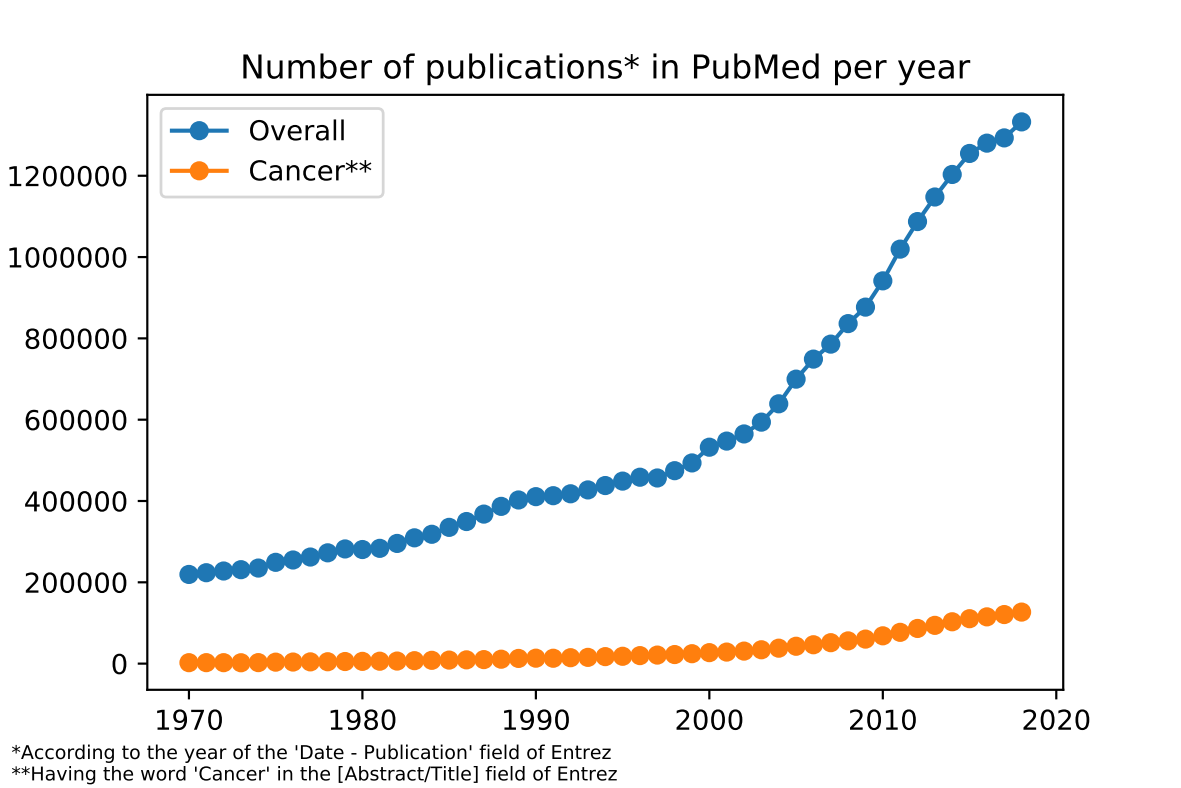
Why Biomedical NLP?

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.



Why Biomedical NLP?

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.



Agenda

- Biomedical NLP 101: Bags of words (30 mins)
- Deep Learning for Biomedical NLP (30 mins)
  - Language Modelling (30 mins)
  - Text Classification (30 mins)
  - Named Entity Recognition (30 mins)
  - Question Answering (30 mins)
  - Integrating NLP into survival models (30 mins)