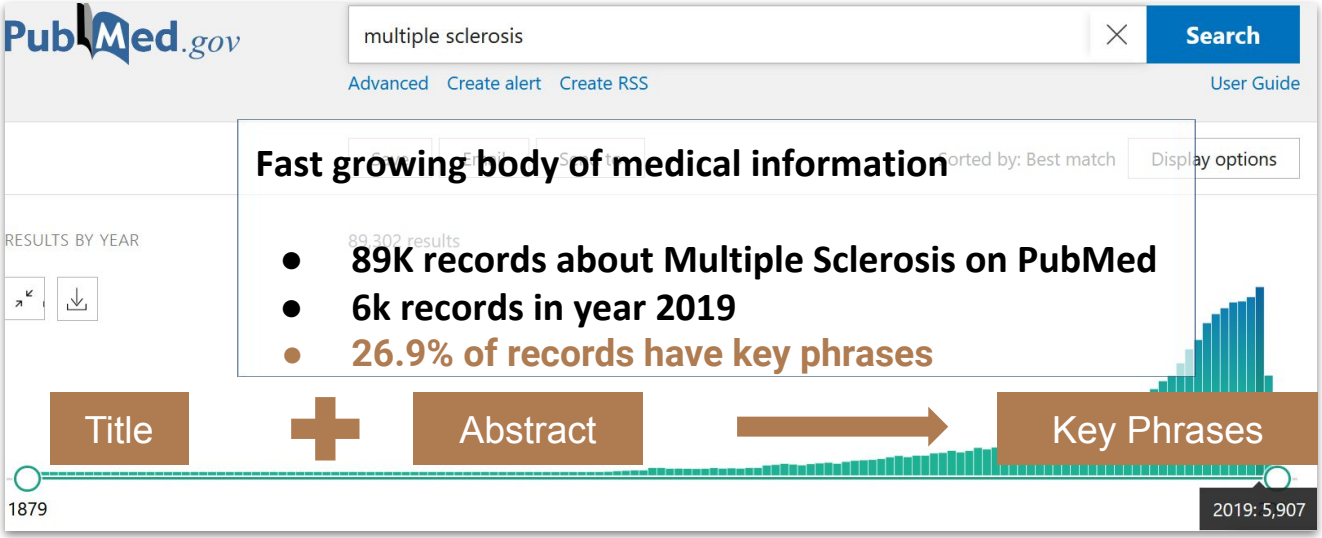
tl;dr
Key Phrase Extractor for Scientific Literature

Pengcheng Ding

Alphabet Health
ADVANCING HEALTHCARE THROUGH SCIENTIFIC COMMUNICATIONS



Key Phrases

Indexing
Summarization
Categorization

Key Phrase Extraction

Extraction: which existing words (in title+abstract) are likely to be part of key phrases

- Tf-idf, TextRank
- Sequence to sequence models: LSTM, LSTM-CRF (Conditional Random Field)
  - Innate relationships among words: embedding layer
  - Long and short term dependencies among the words in the text

Metric: Extraction Rate

Author-provided key phrases:

Forensic Psychiatry; Lyme Disease; Multiple Sclerosis; Propaganda; Psychosis. 60%

Sequence to sequence labelling model

Multiple Sclerosis - A Review. Multiple sclerosis (MS) is the commonest non-traumatic disabling disease to affect young adults … The epidemiology of MS indicates that low serum levels of vitamin D, smoking, childhood obesity and infection with the Epstein-Barr virus are likely to play a role in disease development … There is now the possibility of a diagnosis of 'pre-symptomatic MS' being made … MS epidemiology, potential aetiological factors and pathology are discussed, before moving on to clinical aspects of MS diagnosis and management.

Author provided key phrases
diagnosis; epidemiology; multiple sclerosis

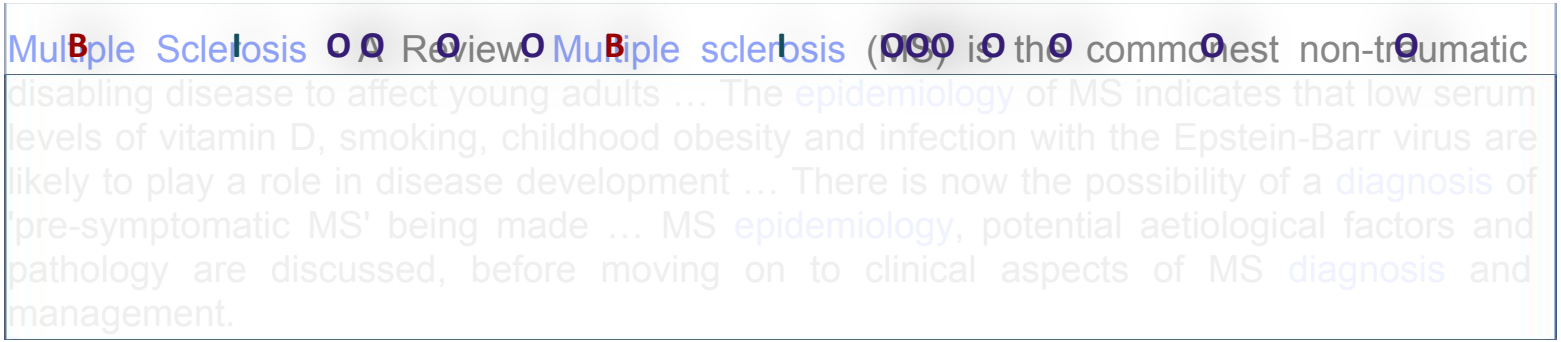
Sequence to sequence labelling model

Multiple Sclerosis - A Review. Multiple sclerosis (MS) is the commonest non-traumatic disabling disease to affect young adults … The epidemiology of MS indicates that low serum levels of vitamin D, smoking, childhood obesity and infection with the Epstein-Barr virus are likely to play a role in disease development … There is now the possibility of a diagnosis of 'pre-symptomatic MS' being made … MS epidemiology, potential aetiological factors and pathology are discussed, before moving on to clinical aspects of MS diagnosis and management.

Author provided key phrases
diagnosis; epidemiology; multiple sclerosis

Sequence to sequence labelling model

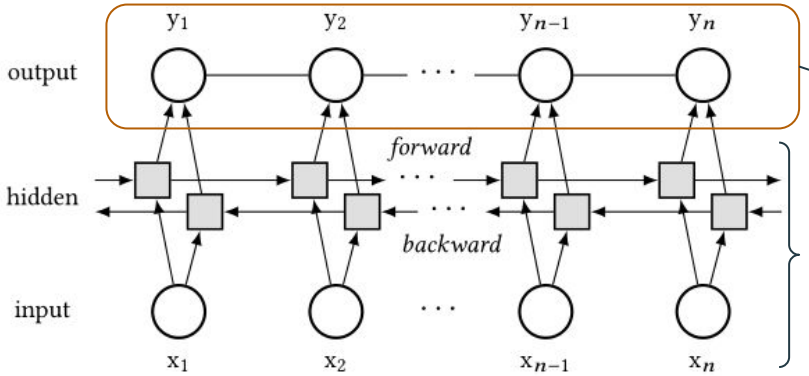
B: Beginning of a key phrase
I: Continuation(inside) of a key phrase
O: Not part of a key phrase



Diagnosis; epidemiology; multiple sclerosis

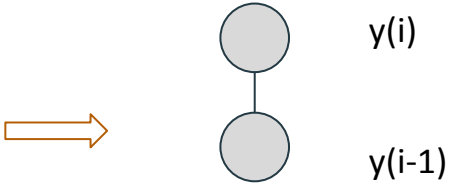
Sequence to sequence labelling model

bi-LSTM and bi-LSTM-CRF (Linear chain conditional random field) models



bi-LSTM

Label-transition matrix (3*3)



Linear Chain CRF

Trained on 11k records

Results reported on validation set (3k)

	bi-LSTM	bi-LSTM-CRF
Accuracy	95.2%	94.7%
F1	42.3%	53.5%
Extraction Rate	39.2%	52.7%

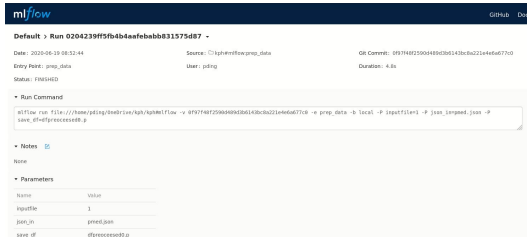
Deliverables(End-to-end Model)

Input → mlflow → JSON

Command line input:

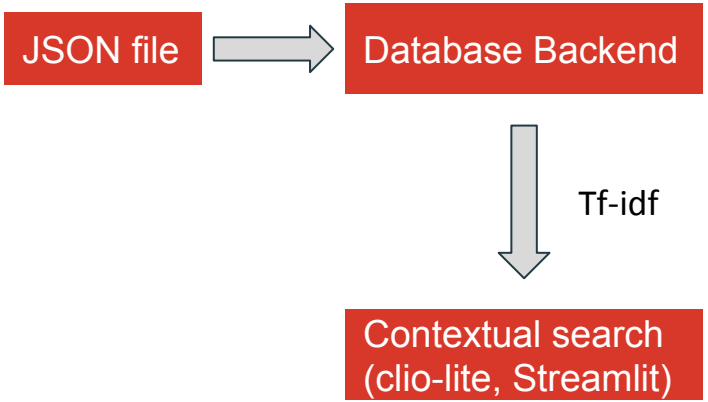
- PubMed search term
- Model parameters
- Module selection
  - Pull records
  - Preprocess
  - Model training
  - Inference

- Tracking
- Monitoring
- Model storage
- Multiple deployment options (dockerized, conda environments)



As input to Power BI

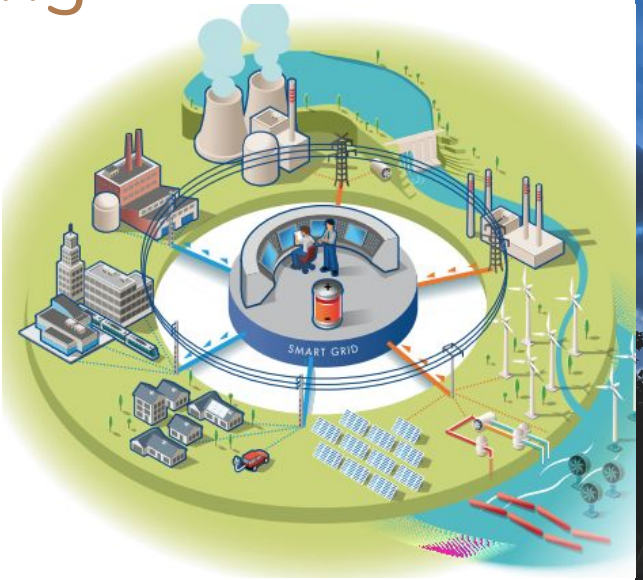
Deliverables (Database)



Results Table

	PMID	title	keywords	pubdate
0	23786735	Utility of the Canadian Treatment Optimization Recommendations (TOR) in MS care.	['multiple sclerosis', 'treatment optimization', 'disease-modifying treatment']	2013
1	24729689	Treatment selection and experience in multiple sclerosis: survey of neurologists.	['disease-modifying therapy', 'multiple sclerosis', 'physician survey', 'treatment adherence', 'treatment satisfaction', 'treatment selection']	2014

Pengcheng Ding



NANJING UNIVERSITY 1902

ÉCOLE POLYTECHNIQUE

IP PARIS

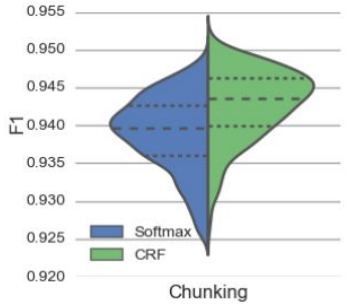
JOHNS HOPKINS
ENVIRONMENT, ENERGY, SUSTAINABILITY & HEALTH INSTITUTE

Variance in the model

Random seed in model training and in train/test split could affect model performance

One instance before API wrapper change:

Extraction rate: LSTM: 49%, LSTM-CRF: 54%



Optimal hyperparameters for deep lstm-networks for sequence labeling tasks
N Reimers, I Gurevych
arXiv preprint arXiv:1707.06799

<pad token>

Padding not automatically recognized

- Two solutions: add confidence to the logits output of LSTM layer on <pad token>s
  - Very slow
- Mask the input to the CRF layer

Context window

Traditional Linear Chain CRF may also include a context window to use x(k-1), x(k), and x(k+1) to compute y(k).

However, no performance increase has been observed after I implemented this with the LSTM model

- Harder to train and to regulate
- bi-LSTM already contained this information

Embedding Layer Regularization

Dropout: any input element

Adjacent elements from next word may still provide relevant information → not successful regularization

Spatial dropout: dropping the same embedding dimension across all words