Building an enterprise Natural Language Search Engine with ElasticSearch and Facebook's DrQA

Louis Baligand, Debmalya Biswas

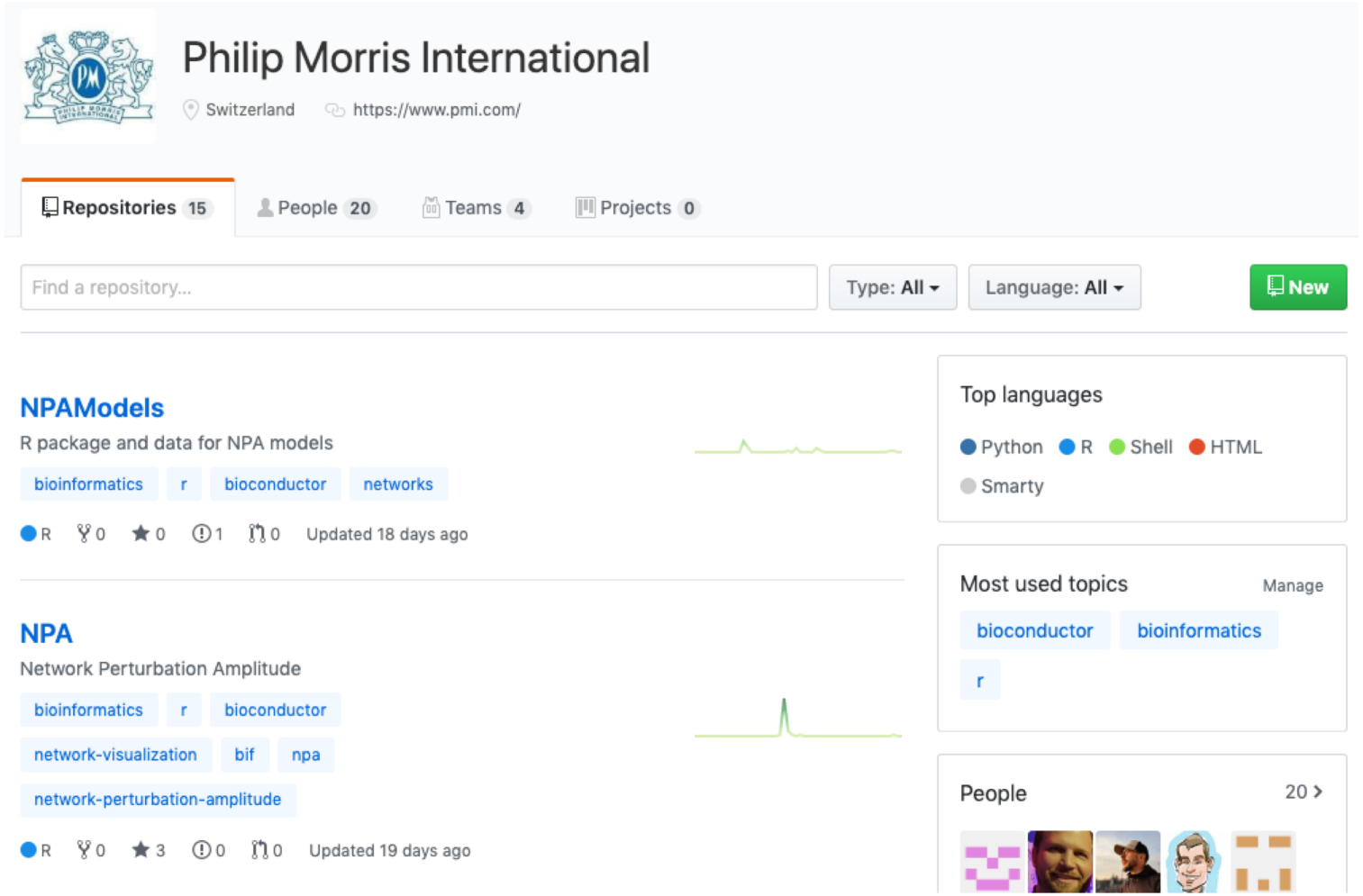
Berlin Buzzwords, 17 June 2019

About

Debmalya Biswas

Louis Baligand

https://github.com/philipmorrisintl



Forrester defines cognitive search and knowledge discovery solutions as

A new generation of enterprise search solutions that employ AI technologies such as natural language processing and machine learning to ingest, understand, organize, and query digital content from multiple data sources.

" The average interaction worker spends [...] nearly 20 percent (of the workweek) looking for internal information."
-MGI Report, 2012.

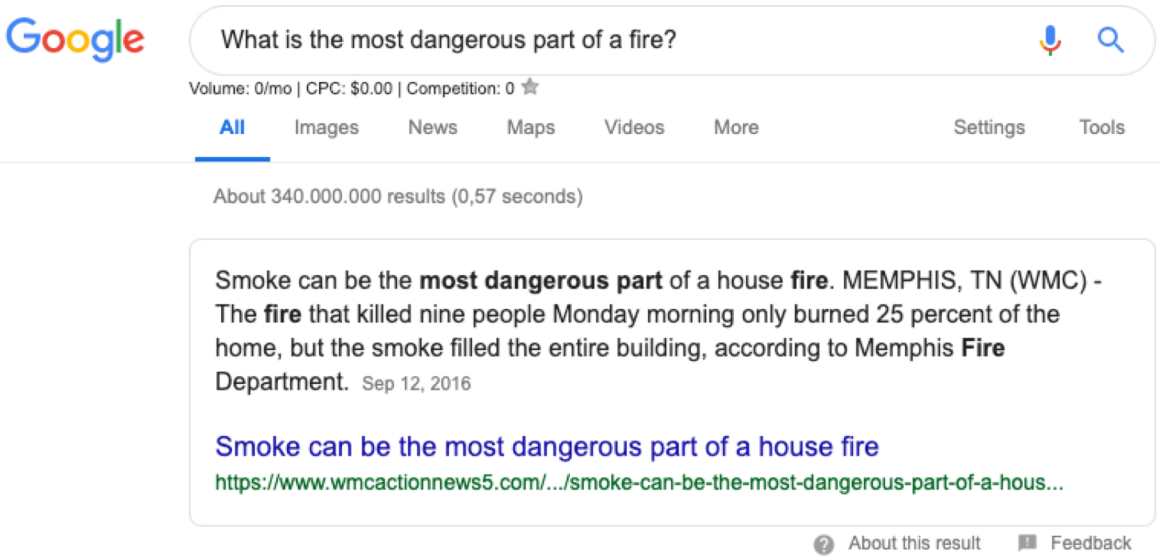
Half (54%) of global information workers said, "My work gets interrupted because I can't find or get access to information I need to complete my tasks" a few times a month or more often. -Forrester Data Global Business Technographics Devices And Security Workforce Survey, 2016.

Enterprise Search vs. Web Search

Enterprise Search	vs.	Web Search
Small amount of content		Enormous amount of content
Employees are the end-users		WWW users
Multiple content types		Single source (web pages)
Limited tagging/metadata management		Large investments in SEO (*)
Role-based content trimming		No visibility restrictions (public pages)
No team in charge of Search Experience		Search xxperience as core business

(*): Search Engine Optimization

Natural Language Search (NLS)



Knowledge Graph

Chatbots and Natural Language Search

(Rules based) FAQs

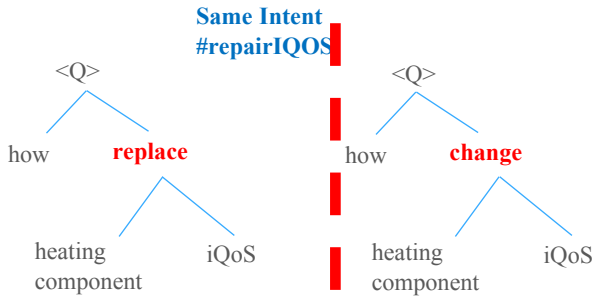
.

- Works only for specific hardcoded questions.
- The only way to scale with respect to question variants, is to extend the knowledgebase by manually adding variants of a question.

"How do I replace the heating component of my iQoS?"
=
"Tell me how to change the heating component of my iQoS"

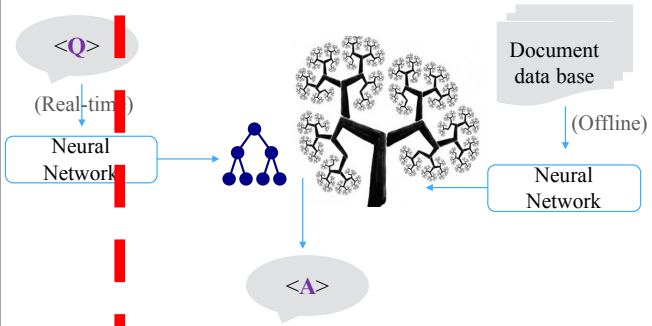
Intent based Chatbots (Statistical Methods)

- Requires Q&A knowledge.
- Able to scale with respect to question variants by applying Statististical Clustering Methods, e.g. tf-idf, Bag-of-Words - to cluster question variants into 'intents'.



Natural Language Search (Neural Networks)

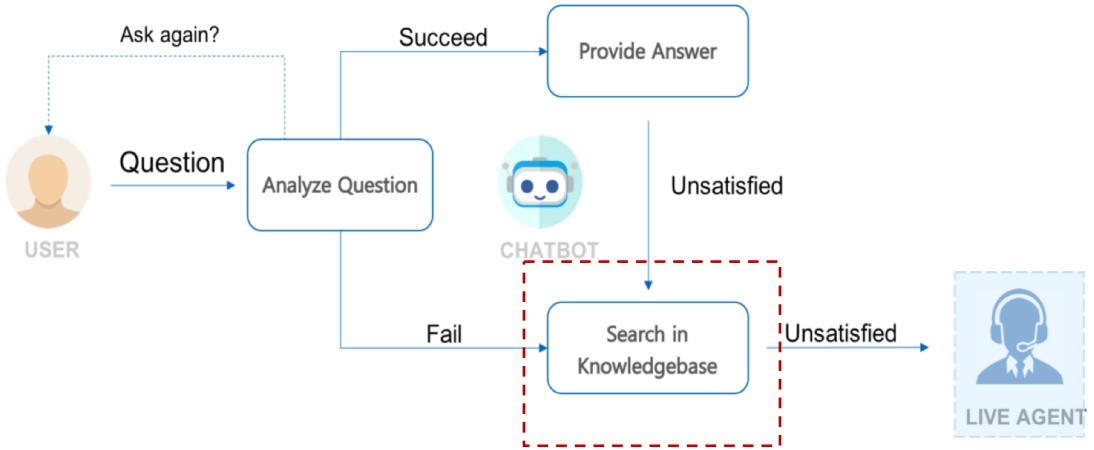
- Works on documents.
- Users can ask any question from the documents.
- Both the documents and questions are passed through the same Neural Network, producing the matching answer.

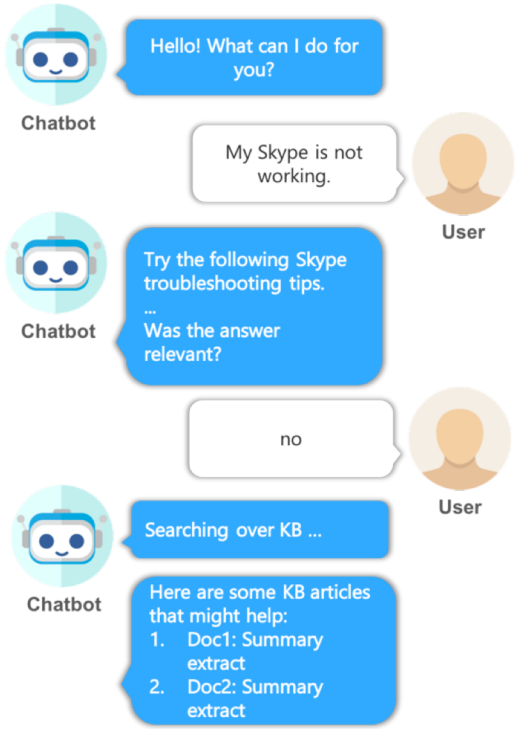


Chatbots and Natural Language Search (2)

3- tier strategy:

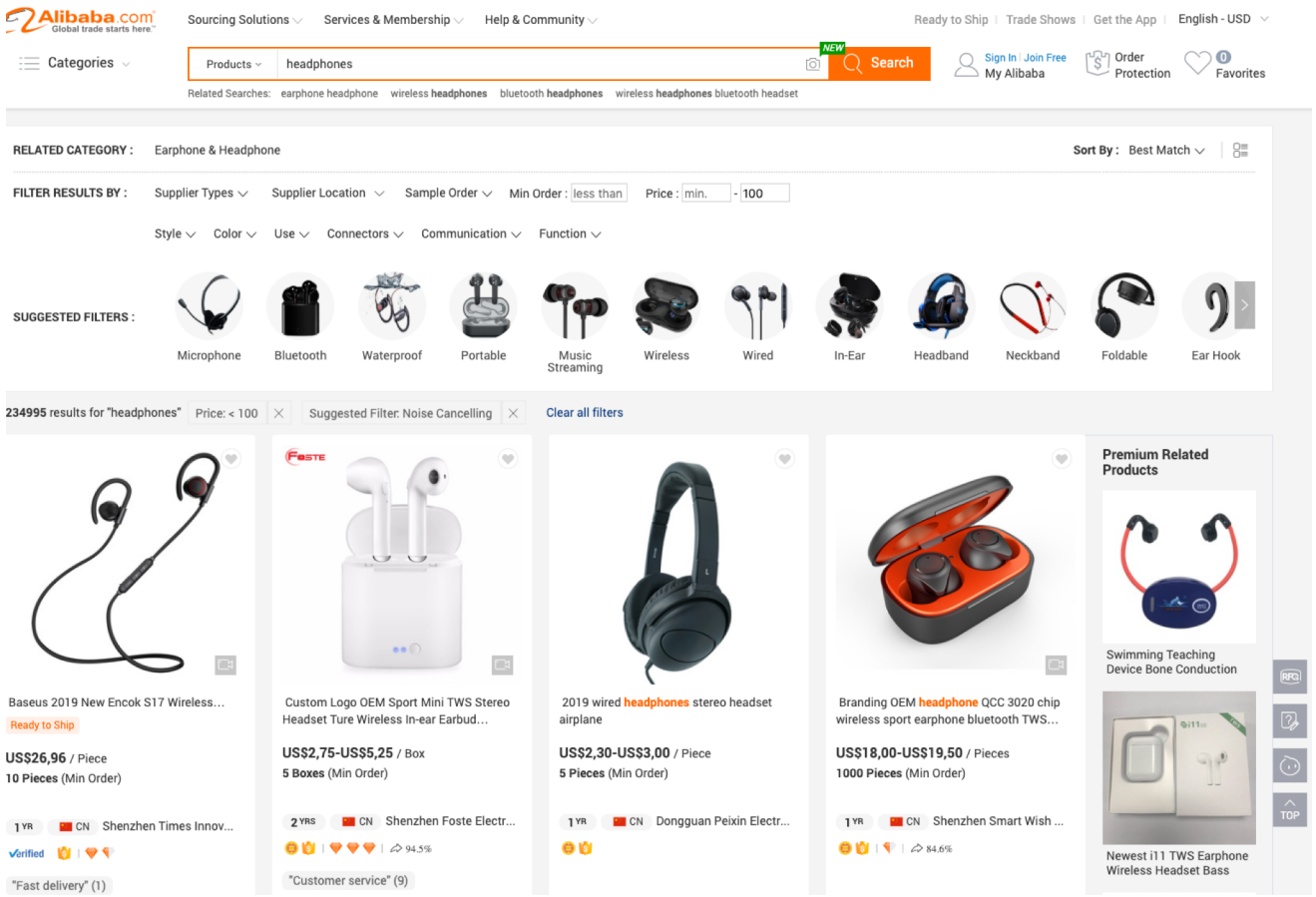
- A Chatbot with its pre-defined Q&A set remains the entry point – think of it as the 1st line of defense.
- If the bot encounters a user query which cannot be mapped to one of its pre-configured intents, it performs a NLS over its KB. This is the 2nd line of defense.
- If the user is not satisfied even with search results, plan for a final handover to a live agent.





Positioning vs e-commerce search

- End-user searching for products (not answer)
- Filter-Oriented
- Rates, Review



Philip Morris' Use case: Operator Trainings

- Hundreds to thousands of operators
- Long manuals with specific terminology
- A 1min downtime of a machine would lead to 20,000 cigarettes unmade
- Typical Full text Search (Boolean search, no relevancy score)
- Document Management System Manually classified
- On-boarding difficulty

Example of fine-grained results

FILE TYPE	COUNT
PDF	798
MS-Office	434
Image	627
Video	57
Zip	73

Q. How many knives are there on the drums?

Type in your question here.

GD121 – AF12 OPERATORS TRAINING MANUAL OPERATION DEFINITION With always changing feeding speed by the tipping paper sliding pallet (31) and the glued tipping paper received from the glue application group (C), would be kept ready for cutting by the tipping paper drum (25). There are fourteen knives which are operating oppositely on the drums (24-25). The glue applied side of the tipping paper which has been fed upward, would be cut by the adjustable knives and would be transported to the next drums for the continuation of the process. The cut tipping papers which has been defined as waste would be transported to the waste pot located at the rear side of the machine. The necessary vacuum at the transporting moment would be provided by a separated fan. The paper should be fed at the same speed with the tipping paper drum (25) and knives during the cutting operation. At the cutting moment, the paper would be held and would be slung on the tipping paper drum vacuum plates for the reason of keeping the paper ready for the next cut. PAGE 250 PHILSA TECHNICAL TRAINING DEPARTMENT PREPARED BY:A.SEYHAN GÜRELLI Technical Trainer

Q: How many knives are there on the drums?

A:

Rank	Answer	
1	fourteen	1. Cigarette Making Equipment\5. GD 121\1.0 Operator\01.GD121-AF12 OPERATORS MANUAL.pdf (page 250 of 357)
2	177	1. Cigarette Making Equipment\5. GD 121\1.0 Operator\01.GD121-AF12 OPERATORS MANUAL.pdf (page 5 of 357)
3	15	1. Cigarette Making Equipment\5. GD 121\1.0 Operator\01.GD121-AF12 OPERATORS MANUAL.pdf (page 121 of 357)

Question Answering?

- Squad Dataset: a reference in Question Answering
- 100,000+ Q&A on Wikipedia articles
- State of the art is beating Human Performance

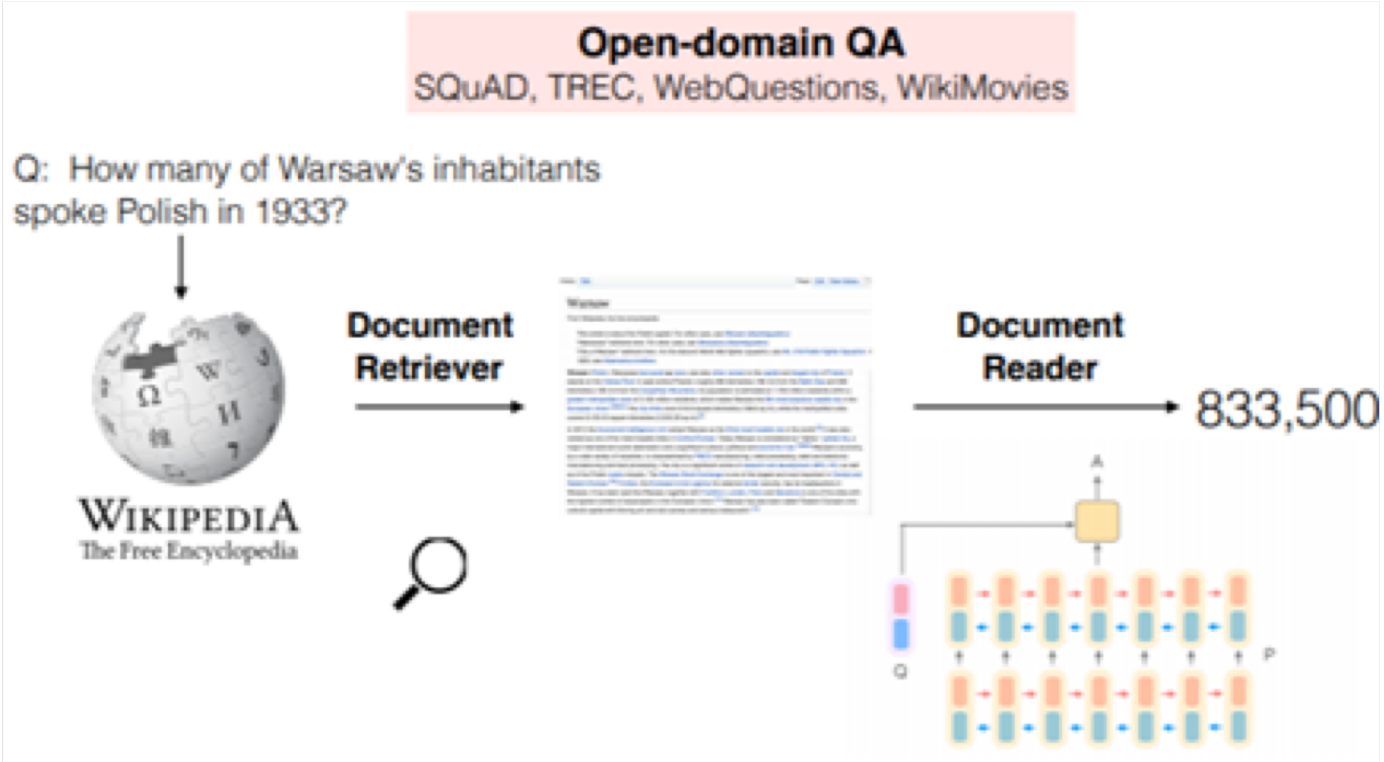
uAD1.1 Leaderboard

re are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	XLNet (single model) *XLNet Team*	**89.898**	**95.080**
2 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
3 May 14, 2019	ATB (single model) *Anonymous*	86.940	92.641

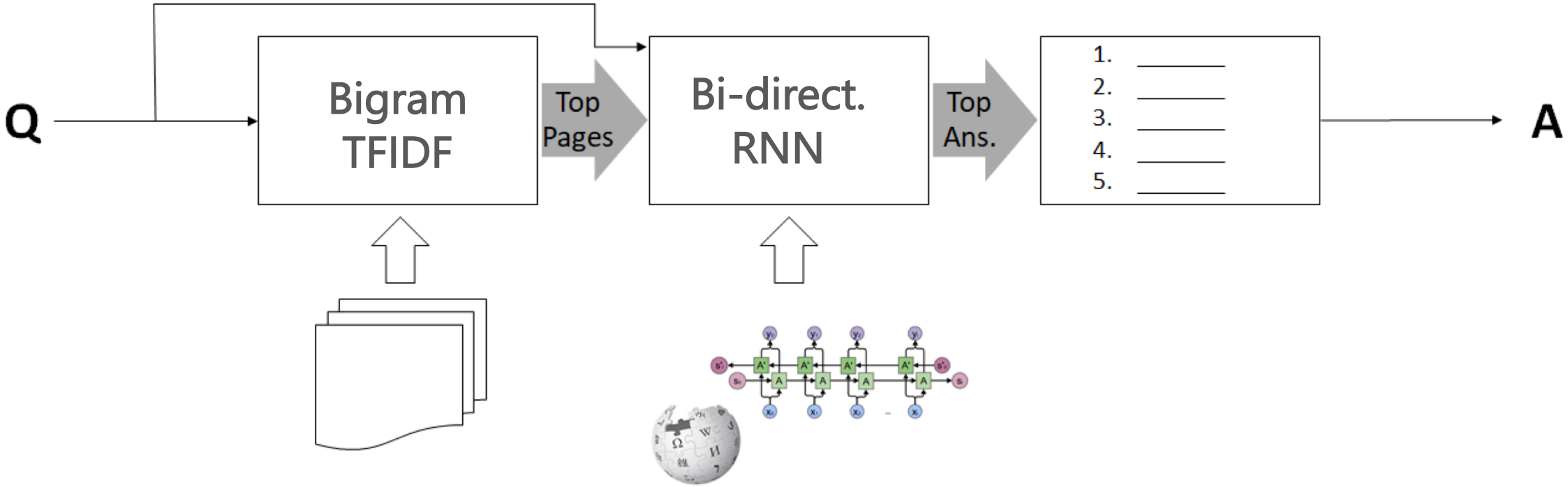
DrQA Overview

- Facebook AI Research, ACL 2017, [Reading Wikipedia to answer Open-Domain Questions](#).
- Open source, BSD License https://github.com/facebookresearch/DrQA
- Pre-trained model available



https://github.com/facebookresearch/DrQA

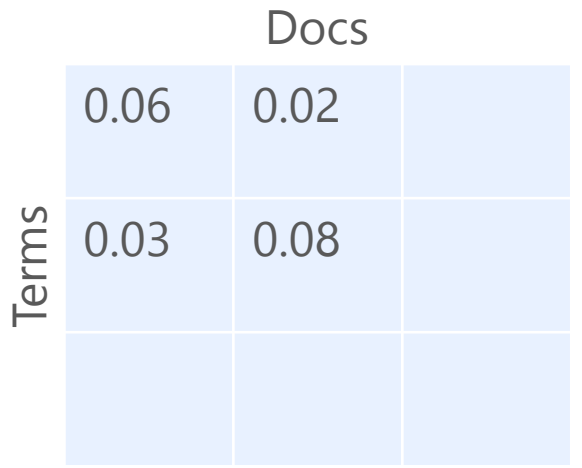
DrQA Overview



DrQA is easy to use on your own corpus!

```
$ python build_db.py /path/to/data /path/to/saved/db.db

$ python build_tfidf.py /path/to/doc/db /path/to/output/dir
```



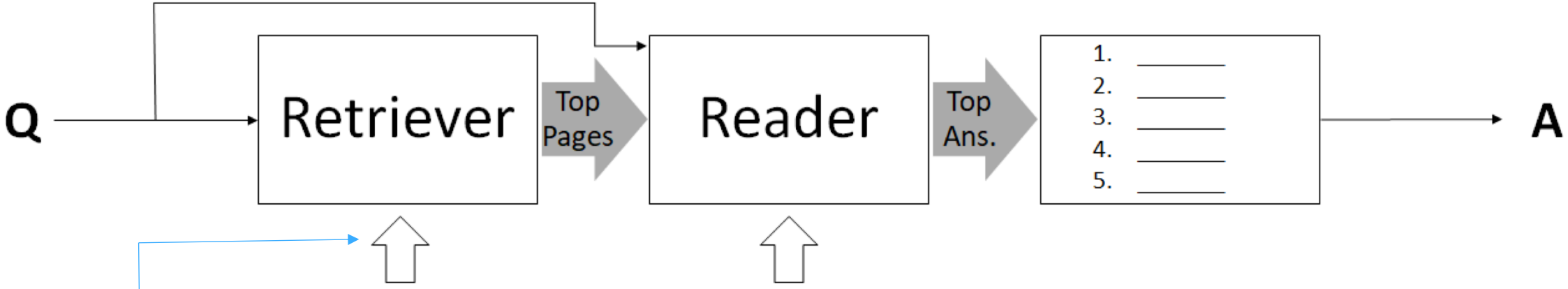
Pre-trained model open sourced

```
$ python interactive.py –reader-model multitask.mdl –retriever-model path/to/tfidf –doc-db
path/to/saved/db.db

>>> process('What is the answer to life, the universe, and everything?’)

Top Predictions:
+------+--------+-----------------------------------------------+--------------+-----------+
| Rank | Answer |                      Doc                      | Answer Score | Doc Score |
+------+--------+-----------------------------------------------+--------------+-----------+
|  1   |   42   | Phrases from The Hitchhiker's Guide to the Galaxy |      47242   |   141.26  |
+------+--------+-----------------------------------------------+--------------+-----------+
```

DrQA to answer Operator's questions?



Apache™ Tika

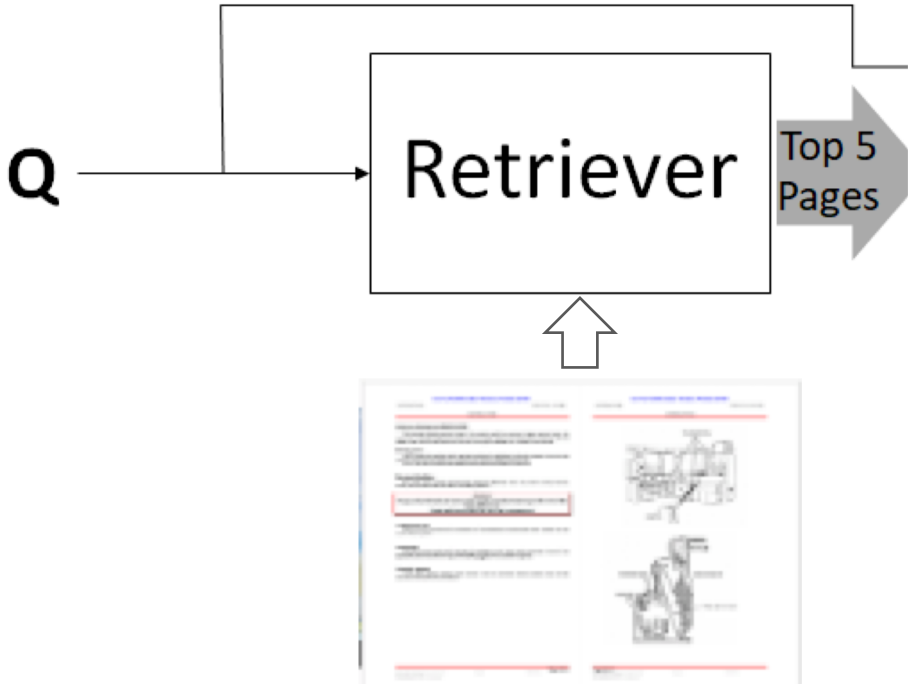
- Java toolkit to extract text + metadata from DOCX, PPT, XLS, PDF, JPEG, etc…
- Apache Software Foundation
- OCR

DrQA to answer Operator's questions?

Model	SQuAD P@5	CuratedTREC P@5	WebQuestions P@5	WikiMovies P@5	Size
TF-IDF model	78.0	87.6	75.0	69.8	~13GB

P@5 here is defined as the % of questions for which the answer segment appears in one of the top 5 documents.

https://github.com/facebookresearch/DrQA



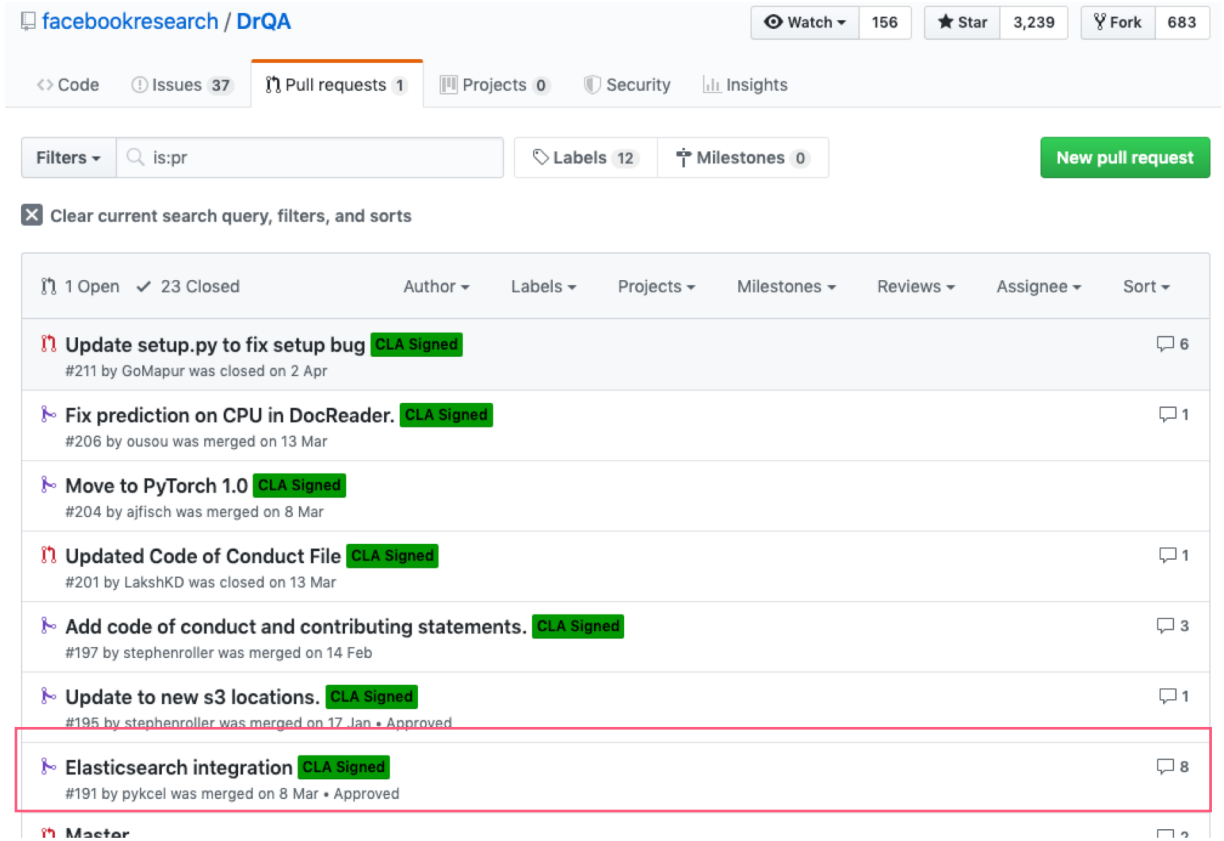
P@5: 76%

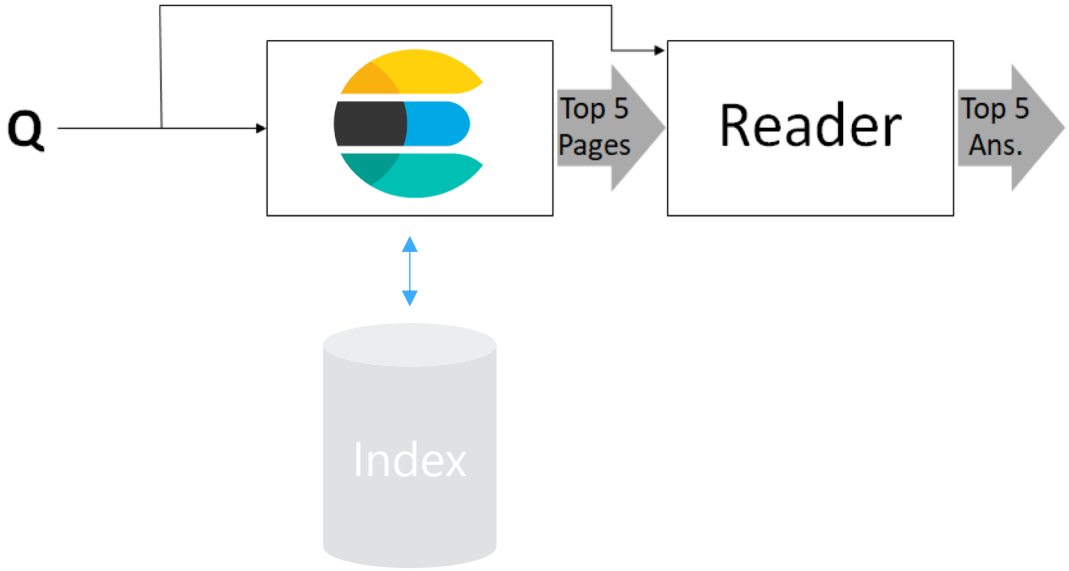
- Not a voice assistant
- End user needs at least ~95%
- Full control on the retriever
- First stage to prioritize

Introducing Elasticsearch

- Open source distributed
- Highly scalable
- RESTful API on top of Lucene capabilities
- Support for Full Text search (best of bread)
- Easy to configure + extend
- Seamlessly manage conflicts
- Active community & popular

Integrating Elasticsearch to DrQA's pipeline





Integrating Elasticsearch to DrQA's pipeline

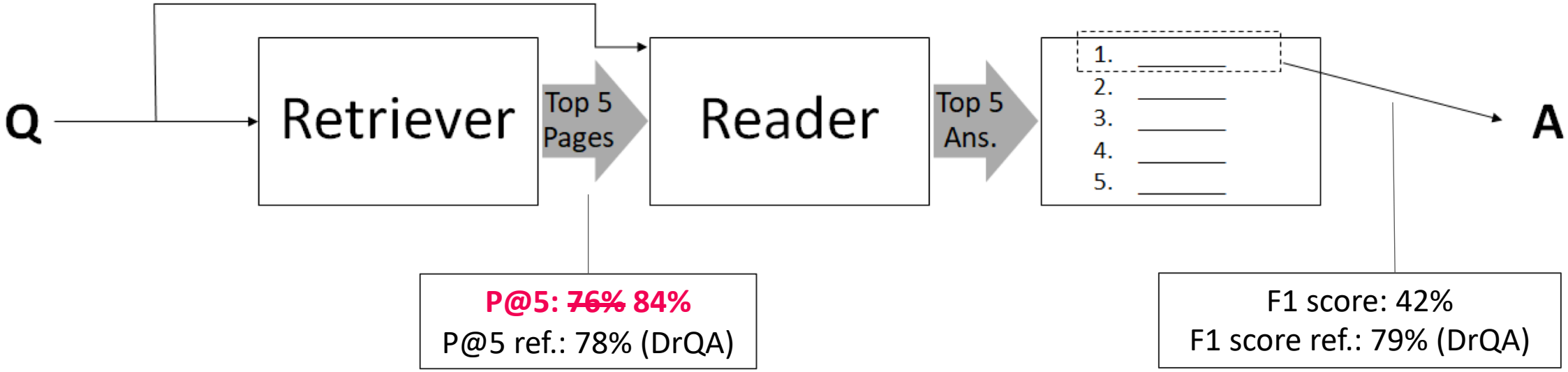
```
>>> from drqa.pipeline import DrQA
>>> from drqa.retriever import ElasticDocRanker
>>> model = DrQA(reader_model='reader_model.mdl',
                 ranker_config={'class':ElasticDocRanker,
                                'options':{'elastic_url':'127.0.0.1:9200',
                                           'elastic_index':'mini',
                                           'elastic_fields':'content',
                                           'elastic_field_doc_name':['file', 'filename'],
                                           'elastic_field_content': 'content'}})
>>> model.process('How the tensioning of the V-belts should be done?')
```

Directly point to your server hosting Elastic

Enable to search in any fields, e.g. uni-grams, bi-grams, title, metadata, etc…


```
29    "_index": "mini",
30    "_type": "doc",
31    "_id": "ae123782374bf93209fg",
32    "_score": 2.34598412931,
33    "_source": {
34      "filename": "manual_machine_example.pdf",
35      "last_modified": "2019-02-29T12:09:02.261+0000",
36      "author": "Louis Baligand",
37      "content": "Follow these steps to do the tensioning of the V-belts:
```

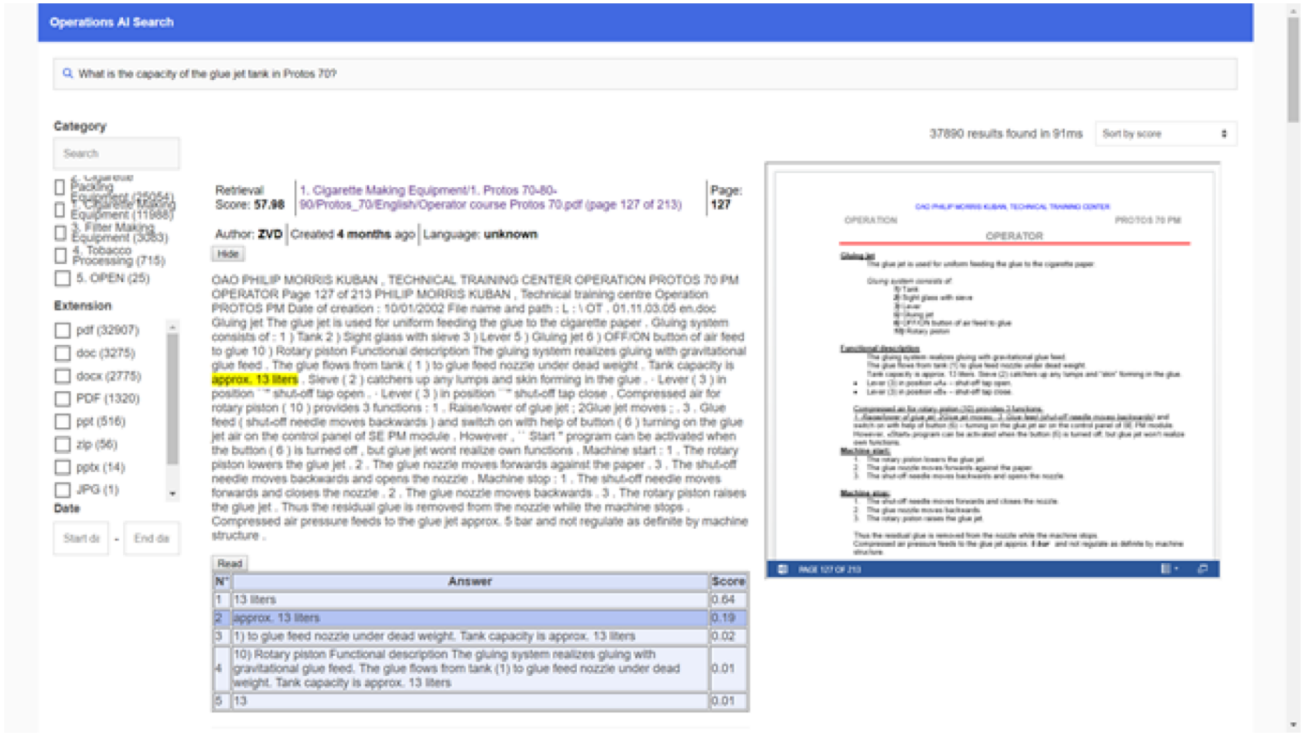
The pipeline performance



- DrQA span +- 10 tokens: **94%** of 1st result contains true answer

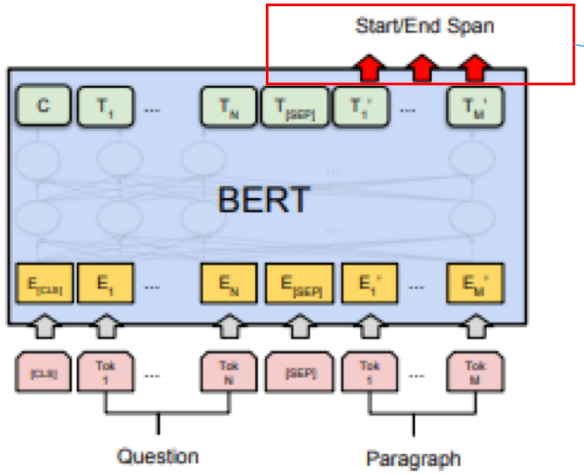
Take aways

- Address pain points by combining Document Retrieval with Question Answering
- If not answered, it will provide much more granular insights of the data
- User elicitation & user experience: a top down approach
- End user does not know what to ask



Future work – Extend pipeline with BERT*

- A general-purpose architecture to train models for multiple NLP tasks (sentiment analysis, etc…)
- State of the art for SQuAD
- Open source, published in Oct. 2018 by Google AI Research
- High memory required: GPU with at least 12GB of RAM (Base model)
- Enable to multi-language queries



(c) Question Answering Tasks:
SQuAD v1.1

- Add one layer to compute P_{start}("token") & P_{end}("token") for each tokens
- Find the best pair by maximizing P_{start}("token1") * P_{end}("token2")

*https://arxiv.org/abs/1810.04805, https://github.com/google-research/bert

Thank you.

INFORMATION SERVICES