Question Answering with Hybrid Data and Models

Sanjay Kamath Ramachandra Rao[1,2]

Supervisors: Brigitte Grau[1], Yue Ma[2]

[1] LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[2] LRI, Université Paris-Sud, CNRS, Université Paris-Saclay, France

Project: ANR GoASQ

February 6, 2020

Plan

1. Introduction
2. State of the art
3. Building Domain-Specific Models
4. Leveraging Semantic Information
5. Conclusion

Question Answering

Question Answering - Introduction

Question Answering

- A research domain dealing with answering questions.
- Natural Language questions on plain text, databases or knowledge bases.
- Factoid and Non-Factoid questions.

Question Answering - Types of Questions

Q: What country are Volvo automobiles made in? (Location - Country)
A: Sweden

Q: How tall is Mount McKinley? (Numerical value - Height)
A: 6,190 m

Q: What currency is used in Ukraine? (Currency)
A: Ukrainian hryvnia

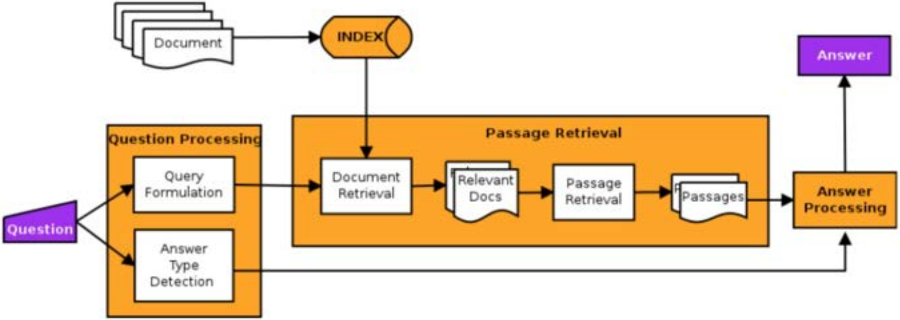
Q: Who played the role of Heisenberg in the series Breaking Bad? (Person)
A: Bryan Cranston

Factoid

Question: Why is ice less dense than water?
Answer Passage: The molecules of water are closer together and constantly moving, whereas the molecules of ice are in a crystal lattice, meaning they're in a rigid formation. When water freezes, the molecules spread out a little more to form the crystal lattice. Since density is mass over volume, and ice has takes up more volume than water, the density of ice is lesser than that of water. Which makes ice float on water.

Non - Factoid

Question Answering Pipeline - The General Approach



- **Question Processing** module analyses questions to detect the Expected Answer Type.
- **Passage retrieval** module uses indexed set of documents to find relevant set of documents and further retrieves a set of relevant paragraphs.
- **Answer Processing** module extracts the answer for the question from the set of relevant paragraphs.

Question Answering Pipeline - Document Retrieval

Question: Who is the President of France?

Document 1: The President of France, officially the President of the French Republic, is the executive head of state of Francein the French Fifth Republic. In French terms, the presidency is the supreme magistracy of the country.
The powers, functions and duties of prior presidential offices, as well as their relation with the prime minister and Government of France, have over time differed with the various constitutional documents since the French Second Republic. The president of the French Republic is also the *ex officio* co-prince of Andorra, grand master of the Legion of Honour and of the National Order of Merit. The officeholder is also honorary proto-canon of the Basilica of St. John Lateran in Rome (although some have rejected the title in the past).
The current president of the French Republic is Emmanuel Macron, who succeeded François Hollande on 14 May 2017.

Document 2: The presidency of France was first publicly proposed during the July Revolution of 1830, when it was offered to the Marquis de Lafayette. He demurred in favour of Prince Louis Phillipe, who became King of the French. Eighteen years later, during the opening phases of the Second Republic, the title was created for a popularly elected head of state, the first of whom was Louis-Napoléon Bonaparte, nephew of Emperor Napoleon. Bonaparte served in that role until he staged an auto coup against the republic, proclaiming himself Napoleon III, Emperor of the French.

Document 3: Under the Third Republic and Fourth Republic, which were parliamentary systems, the office of President of the Republic was a largely ceremonial and powerless one. The Constitution of the Fifth Republic greatly increased the President's powers. A 1962 referendum changed the constitution, so that the president would be directly elected by universal suffrage and not by the Parliament. In 2000, a referendum shortened the presidential term from seven years to five years. A maximum of two consecutive terms was imposed after the 2008 constitutional reform.

QA Pipeline - Paragraph/Sentence Selection

Question: Who is the President of France?

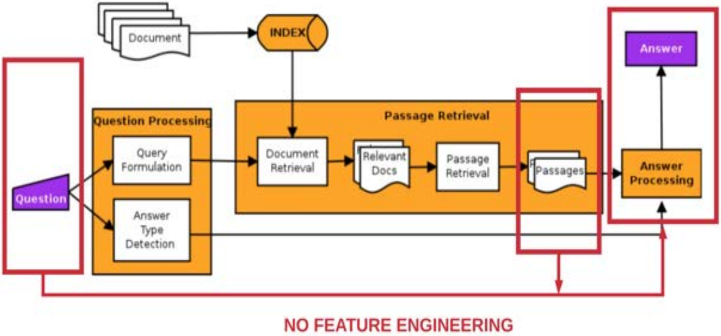
Paragraph 1: The President of France, officially the President of the French Republic, is the executive head of state of Francein the French Fifth Republic. In French terms, the presidency is the supreme magistracy of the country.
Paragraph 2: The powers, functions and duties of prior presidential offices, as well as their relation with the prime minister & Government of France, have over time differed with the various constitutional documents since the French Second Republic. The president of the French Republic is also the *ex officio* co-prince of Andorra, grand master of the Legion of Honour and of the National Order of Merit.
Paragraph 3: The officeholder is also honorary proto-canon of the Basilica of St. John Lateran in Rome (although some have rejected the title in the past).
The current president of the French Republic is Emmanuel Macron, who succeeded François Hollande on 14 May 2017.
Paragraph 4: The presidency of France was first publicly proposed during the July Revolution of 1830, when it was offered to the Marquis de Lafayette. He demurred in favour of Prince Louis Phillipe, who became King of the French.
Paragraph 5: Eighteen years later, during the opening phases of the Second Republic, the title was created for a popularly elected head of state, the first of whom was Louis-Napoléon Bonaparte, nephew of Emperor Napoleon. Bonaparte served in that role until he staged an auto coup against the republic, proclaiming himself Napoleon III, Emperor of the French.
Paragraph 6: Under the Third Republic and Fourth Republic, which were parliamentary systems, the office of President of the Republic was a largely ceremonial and powerless one. The Constitution of the Fifth Republic greatly increased the President's powers.
Paragraph 7: A 1962 referendum changed the constitution, so that the president would be directly elected by universal suffrage and not by the Parliament. In 2000, a referendum shortened the presidential term from seven years to five years. A maximum of two consecutive terms was imposed after the 2008 constitutional reform.

QA Pipeline - Reading Comprehension

Question: Who is the President of France?

Paragraph 3: The officeholder is also honorary proto-canon of the Basilica of St. John Lateran in Rome (although some have rejected the title in the past).
The current president of the French Republic is Emmanuel Macron, who succeeded François Hollande on 14 May 2017.

Deep Learning based Question Answering Pipeline



- Individual modules are based on deep learning based algorithms.
- Increase in interests to build end-to-end models for individual modules.
- Some assumptions on the data can be seen while using deep learning.

Hurdles using Deep Learning models for QA

- Having enough data (Size)
- Having the right kind of labelled data (Suitable type)
- Building an end-to-end model which does everything (Complexity)
- Generalizing the model to work on all QA tasks (Generalization)

Size and Type of the data

Size

- How large is large enough to use a deep learning algorithm?
- Do we always need a large scale data in a specific domain?
- Can similar datasets from other domains be useful?

Type

- Deep learning based approaches mainly focus on building end-to-end models.
- How can we use semantic features effectively along with neural network models?
- Are synthetic datasets and human annotated datasets comparable?

Complexity and Generalization

Complexity of the model

- Are complex model always performing better than simple ones?
- How to choose a good model to experiment on a new dataset?
- How to choose the required hardware needed for experiments? (Number of GPUs or TPUs required)

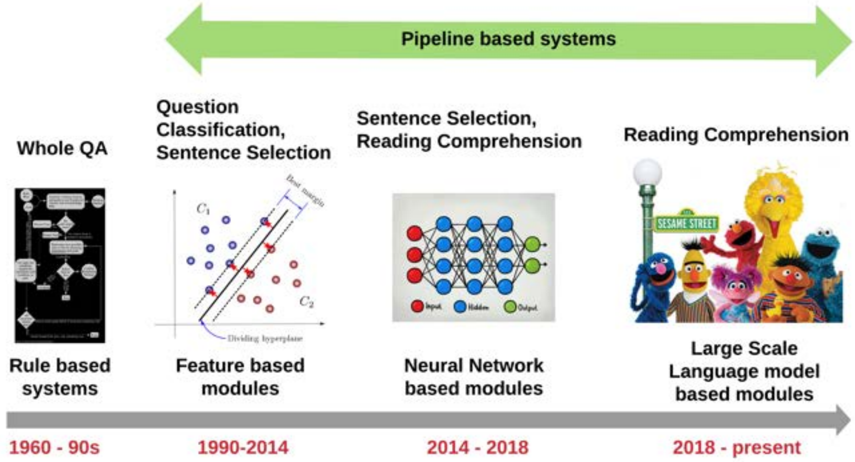
Generalization

- Does one model performing better on a dataset perform similarly on others?
- Does it generalize across different data domains?

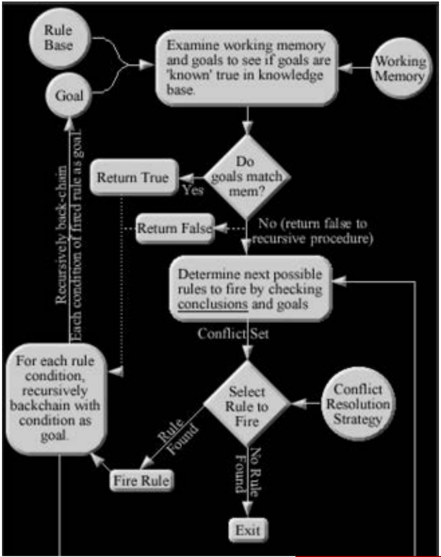
Plan

1. Introduction
2. State of the art
3. Building Domain-Specific Models
4. Leveraging Semantic Information
5. Conclusion

Timeline of different QA system approaches



Rule based systems (1960-1990)



Rule based expert systems

- Hard coded rules by experts.
- Term matching module triggers rules and applies actions.
- BASEBALL (1961), LUNAR (1973), SYNTHEX, LIFER, and PLANES were some of the systems built.

Rule based expert systems - Limitations

- Hard to create rules.
- Extensive amount of human work is required.
- Systems are not robust.
- Not easy to adapt for expert domains.

Pipeline based systems

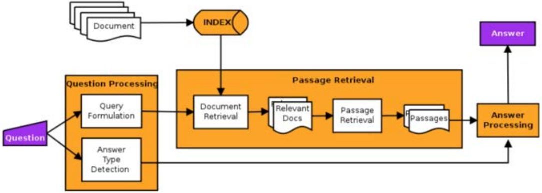
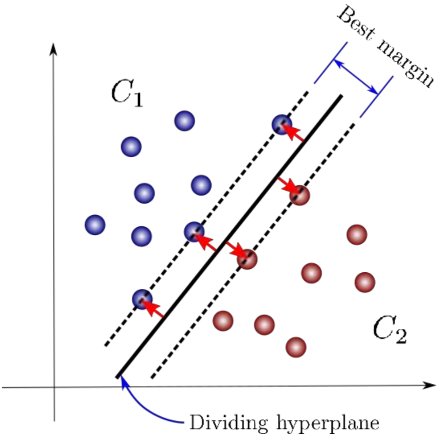


Figure: A typical QA pipeline

Trec QA task

- Trec Question Answering task [Voorhees et al., 2000] since 1999 gave rise to several works which followed pipeline based systems.

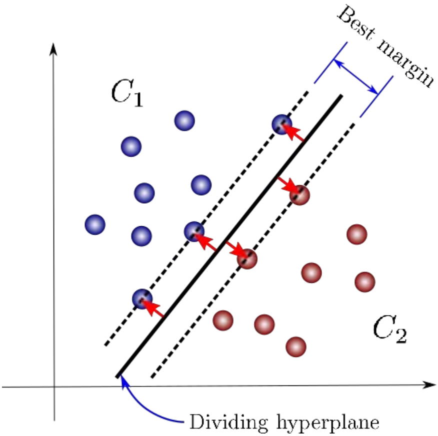
Features based Machine Learning systems (1990-2014)



Machine Learning modules

- Modules rely heavily on input features.
- Individual modules built using ML models on different objectives.
- Question Classification using rules - (Moldovan et al., 1999; Hermjakob, 2001; Radev et al., 2002; Ferret et al., 2001) and ML models - (Hermjakob, 2001).
- Document retrieval uses IR methods.
- Answer processing using Dependency Trees - (Hovy et al., 2002, Punyakanok et al., 2004; Cui et al., 2005).
- Tree edit distances, feature extraction using dependency trees and relations, were used.

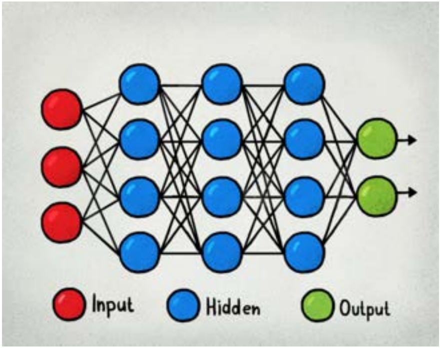
Features based Machine Learning systems (1990-2014)



Machine Learning modules - Limitations

- System performance depends mainly on input features.
- Several NLP tools are used for extracting those features.
- Domain expertise is required for feature extraction.
- NLP tools used for pre-processing may contribute to error propagation.

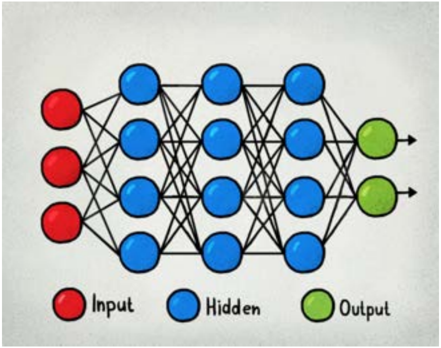
Neural Networks based systems (2014-2018)



Deep Learning modules

- Goal is to avoid Feature Engineering.
- Pioneered with Answer Sentence Selection task using CNN - (L. Yu et al., 2014)
- Several models used RNNs - (Moschitti et al., 2014, He et al., 2015, Yin et al., 2016, Rao et al., 2016)
- SQUAD dataset - (Rajpurkar et al., 2016) triggered lot of work for Reading Comprehension task.

Neural Networks based systems (2014-2018)

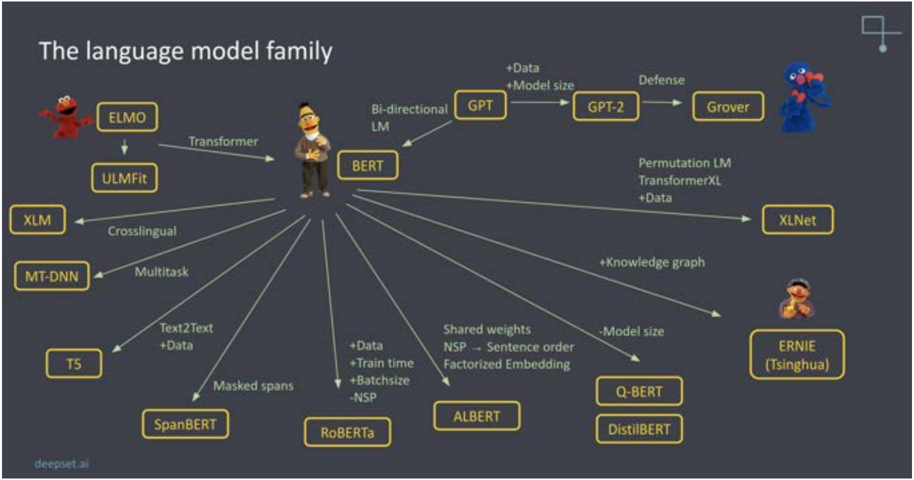


Deep Learning modules - Limitations

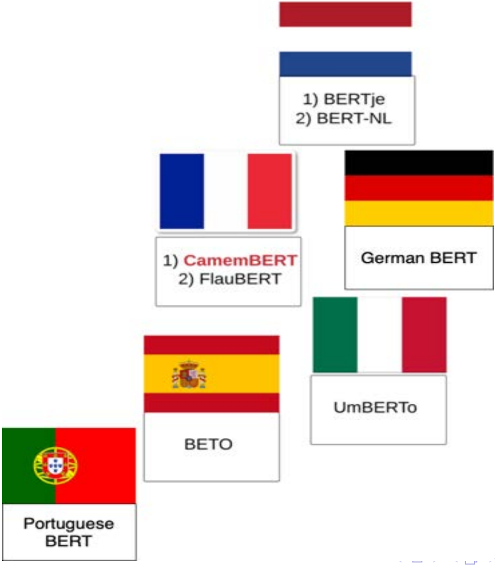
- Choosing right architecture, hyper parameters, optimizer etc. is very important.
- Feature Engineering $\rightarrow \rightarrow$ Architecture Engineering

Large Scale Language Models (2018 - Present)

Large Scale Language Models (2018 - Present)



Large Scale Language Models (2018 - Present)



Large Scale Language Models (2018 - Present)

Large scale LMs for NLP

- Large models trained on large scale datasets.
- ULMFit (Howard et al., 2018) and ELMO (Peters et al., 2017, Peters et al., 2018) used LSTM as units.
- BERT (Devlin et al., 2018) used Bidirectional Transformers.
- BERT and other variants are the current state of the art in several QA tasks.

Research Objectives

1. Building models which work both on small scale and large scale datasets without affecting performance.
2. Leveraging the structured and semantic knowledge effectively into the State of The Art (SOTA) question answering models to improve performance.

Plan

1 Introduction

2 State of the art

3 Building Domain-Specific Models

4 Leveraging Semantic Information

5 Conclusion

Biomedical Domain Data

Biomedical Question Answering task - BIOASQ

- A Question and some paragraphs are provided as input and the system must return Answers.

Question: Which calcium channels does ethosuximide target?
Answer: T-type calcium channels

Snippets:

- Theta rhythms remained disrupted during a subsequent week of withdrawal but were restored with the T-type channel blocker ethosuximide.
- Given evidence that chemotherapy-induced neuropathic pain is blocked by ethosuximide, known to block T-type calcium channels, we examined if more selective T-type calcium channel blockers....
- The Ca(v)3.2 channel is sensitive to ethosuximide, amlodipine and amiloride.

Datasets

	Datasets	Train	Dev	Test
	BIOASQ 4b	427	59	161
Small Scale	BIOASQ 5b	544	75	150
	BIOASQ 6b	685	94	161
Large Scale	SQUAD v1.0	87,599	10,570	9,533
	QUASAR-T	37,012	3,000	3,000

Figure: Large scale and Small scale datasets comparison

Deep Learning and Question Answering

Questions commonly asked

- Do deep learning models always require large scale data?
- With less data, can we not use deep learning methods?
- Can any type of data be used for deep learning?

Our research context

- Using deep learning models on small scale domain specific datasets.
- Using the models built for open domain data towards domain specific data.

QA Task - Reading Comprehension

Question: Which NFL team represented the AFC at Super Bowl 50?

Answer Passage: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24-10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Answer: Denver Broncos. Start Offset: 177

Assumption: Correct Paragraph containing the answer is provided.

BIOASQ Data

Question: Which calcium channels does ethosuximide target?
Answer: T-type calcium channels

Snippets:

- Theta rhythms remained disrupted during a subsequent week of withdrawal but were restored with the T-type channel blocker ethosuximide.
- Given evidence that chemotherapy-induced neuropathic pain is blocked by ethosuximide, known to block T-type calcium channels, we examined if more selective T-type calcium channel blockers....
- The Ca(v)3.2 channel is sensitive to ethosuximide, amlodipine and amiloride.

Figure: Some snippets contain answers and some do not.

Reading Comprehension Format

- Each paragraph containing an answer is considered as a separate QA pair.
- Other non answer containing paragraphs are discarded.

Datasets

Dataset	Gold Questions	RC eligible
Bioasq 4	486	**321 (66%)**
Bioasq 5	619	**428 (69.1%)**
Bioasq 6	779	**543 (69.7%)**
SQUAD v1.0	1,07,702	1,07,702 (100%)

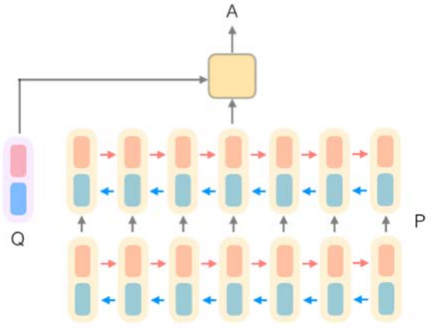
Figure: Datasets used in our experiments which are suitable for Reading Comprehension (RC) setting, as done by [Weissenborn et al., 2018, Lee et al., 2019]

Choosing a Simple Model

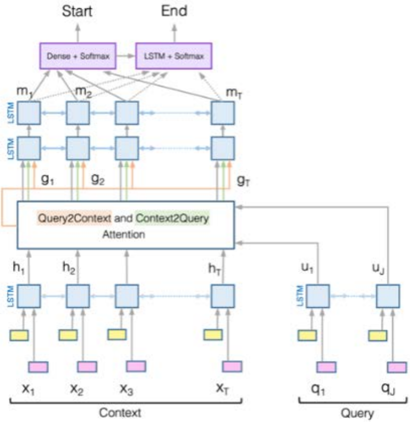
HOW DO WE CHOOSE A SIMPLE MODEL?

- Are all models using LSTMs (comparably similar architecture) the same?
- Can we choose the model which fetches the best scores on a task dataset?
- GPU training time == €€€€!!!
  Is there a tradeoff between training time and accuracy to be considered?
- How does model complexity (more parameters == more complex) affect the training time?

Choosing a Simple Model



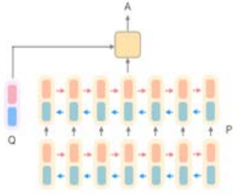
DRQA model

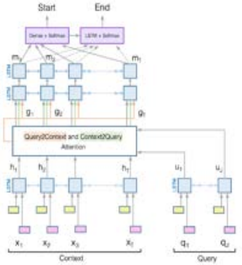


BIDAF model

Figure: DRQA by [Chen et al., 2017] and BIDAF by [Seo et al., 2016]

Choosing a Simple Model





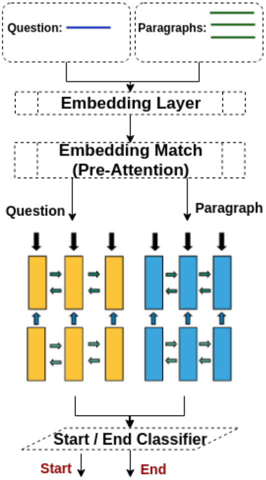
DRQA

- Training time: ∼4 hours on a single GPU
- Exact Match score on SQUAD dataset: 69.5%
- Simple model compared to BIDAF
- Published in March 2017

BIDAF

- Training time: ∼20 hours a single GPU
- Exact Match score on SQUAD dataset: 67.7%
- Complex model compared to DRQA
- Published in Nov. 2016

DRQA



DRQA

- Question (Q) and Paragraphs (P) both are encoded with GLOVE vectors.
- An attention mechanism is used to map embeddings between Q and P.
- $$\mathcal{F}align(p_i) = \Sigma_j a_{i,j} E(q_j) \qquad (1)$$
  Where $a_{i,j}$ is,
  $$a_{i,j} = \frac{exp\,(\alpha(E(s_i)) \cdot \alpha(E(q_j))}{\Sigma_{j'}\; exp(\alpha(E(s_i)) \cdot \alpha(E(q_{j'}))} \qquad (2)$$
- Bi-LSTMS are used individually and are connected to two separate bilinear classifiers for Start and End predictions.
- $$P_{\text{start}}(i) \propto \exp\left(\mathbf{p}_i \mathbf{W}_s \mathbf{q}\right) \qquad (3)$$
  $$P_{\text{end}}(i) \propto \exp\left(\mathbf{p}_i \mathbf{W}_e \mathbf{q}\right) \qquad (4)$$
- Model and implementation by [Chen et al., 2017]

Domain Adaptation using DRQA

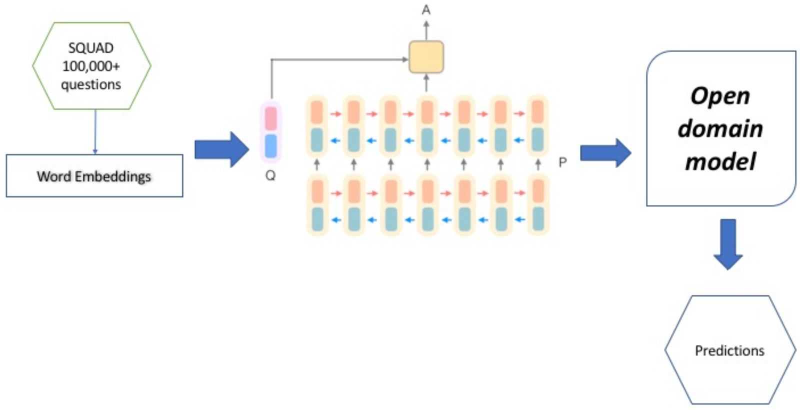


Figure: From open domain towards biomedical domain [Weissenborn et al., 2017]

Domain Adaptation using DRQA

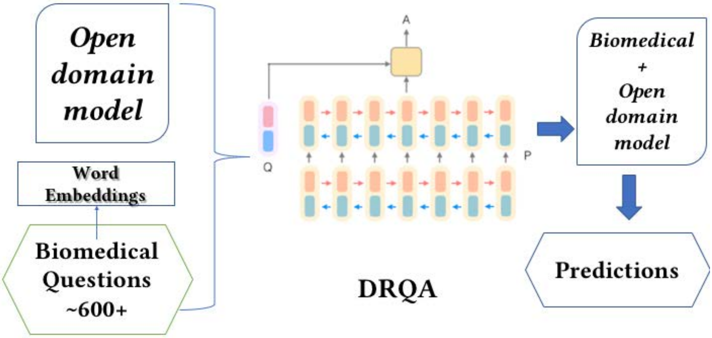
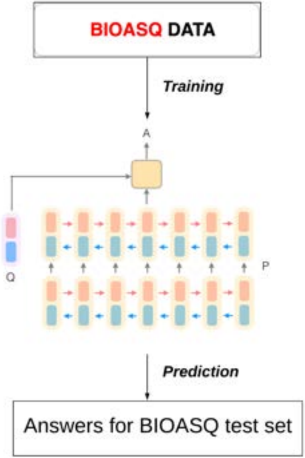
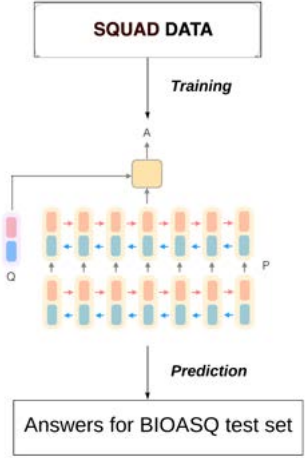


Figure: From open domain towards biomedical domain [Weissenborn et al., 2017]

Pre-training and Finetuning

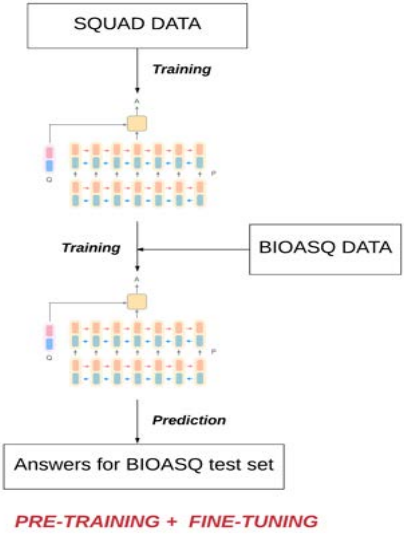




NO PRE-TRAINING

NO FINE-TUNING

Pre-training and Finetuning



Evaluation in BIOASQ

- Automatic evaluation performed on system results using the official evaluation scripts[1].
- Following evaluation measures are computed:
  - → Strict accuracy - rate of exact matching strings of gold standard answers on top 1 prediction.
  - → Lenient accuracy - rate of exact matching strings of gold standard answers in top 5 predictions.
  - → MRR - mean reciprocal rank computed on the top 5 predictions.

[1] https://github.com/BioASQ/Evaluation-Measures

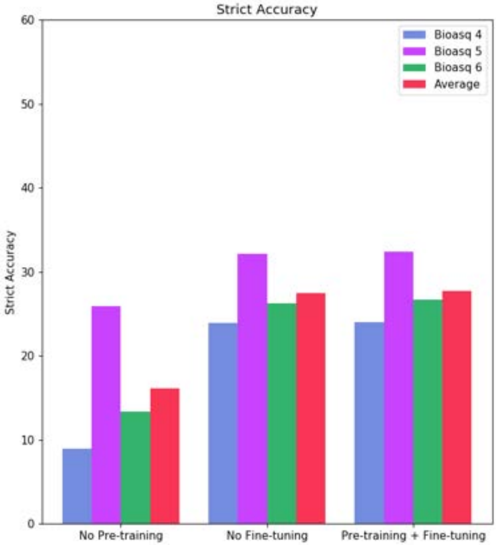


Figure: A study to show the importance of domain adaptation - [Kamath et al., 2019]

Reading Comprehension vs Open QA

Question: Which NFL team represented the AFC at Super Bowl 50?

Answer Passage: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24-10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Answer: Denver Broncos. Start Offset: 177

Q: What's the capital of Ireland?
P_1: As the capital of Ireland, Dublin is...
P_2: Ireland is an island in the North Atlantic...
P_3: Dublin is the capital of Ireland. Besides, Ottawa is one of famous tourist cities in Ireland and ...

BIOASQ Data (Recap)

Question: Which calcium channels does ethosuximide target?
Answer: T-type calcium channels

Snippets:

- Theta rhythms remained disrupted during a subsequent week of withdrawal but were restored with the T-type channel blocker ethosuximide.
- Given evidence that chemotherapy-induced neuropathic pain is blocked by ethosuximide, known to block T-type calcium channels, we examined if more selective T-type calcium channel blockers....
- The Ca(v)3.2 channel is sensitive to ethosuximide, amlodipine and amiloride.

Figure: Some snippets contain answers and some do not.

Reading Comprehension Format

- All paragraphs are considered in the OpenQA model setting.

OPEN-QA model - PSPR

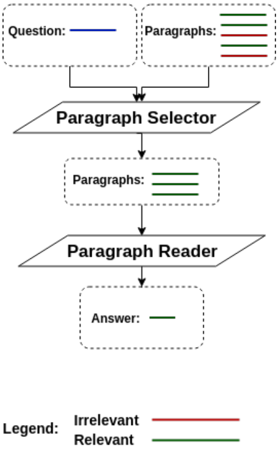


Figure: OpenQA model by [Lin et al., 2018]

OPEN-QA with DRQA

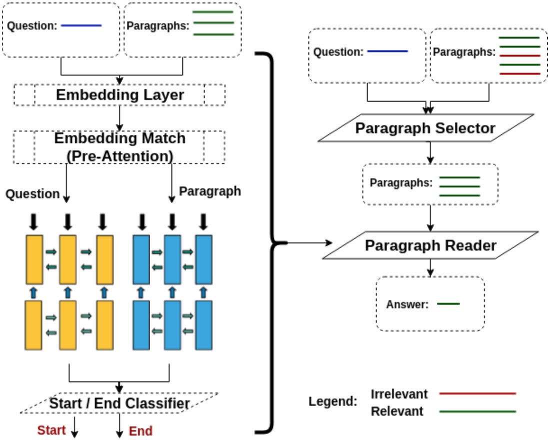
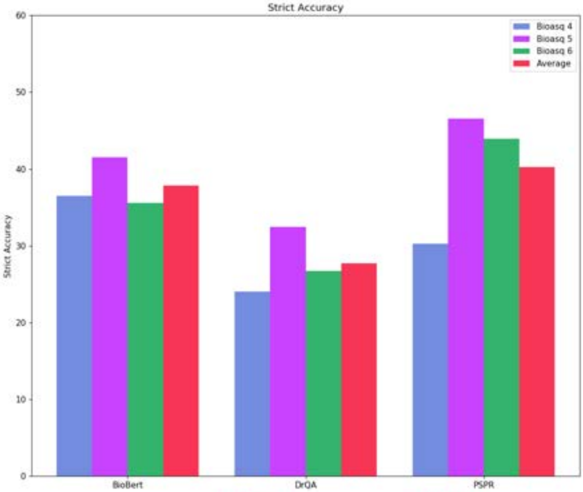


Figure: OpenQA model by [Lin et al., 2018]

Training Times (Approx):

- Biobert:
  Bert - 4 days on 1 TPU pod,
  Biobert - 23 days on 8 GPUs
  and 20 minutes on 1 GPU
  (Fine-tuning)
- DrQA - 4 hours on 1 GPU
- PSPR - 8 hours on 1 GPU



Biobert by [Lee et al., 2019]　　(Ours) DrQA by [Chen et al., 2017]　　(Ours) PSPR by [Lin et al., 2018]

BERT - New SOTA on several QA tasks

BERT - New SOTA on several QA tasks

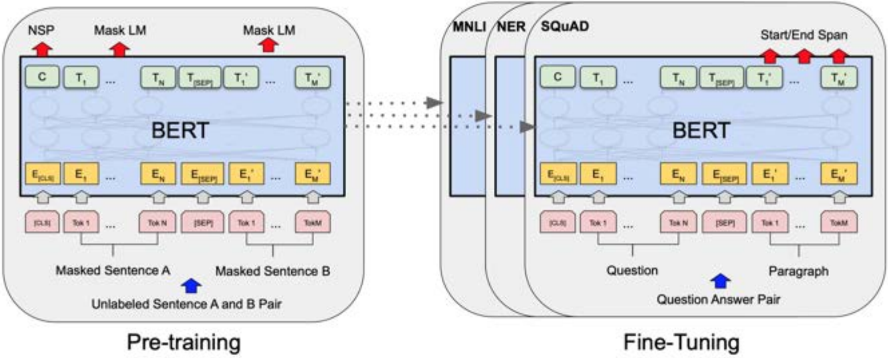


Figure: BERT model by [Devlin et al., 2019]

BIOBERT

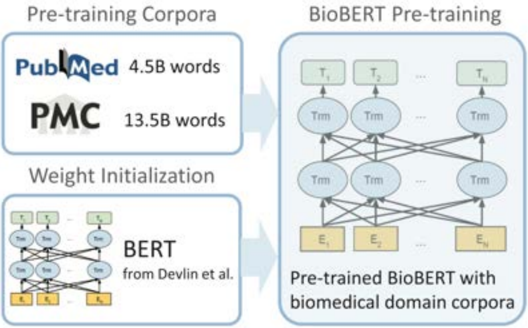


Figure: BIOBERT model by [Lee et al., 2019]

BIOBERT in BIOASQ

- Applied BIOBERT on BIOASQ task data but on document level text snippets.
- Pre-trained with SQUAD data and fine-tuned with BIOASQ data.

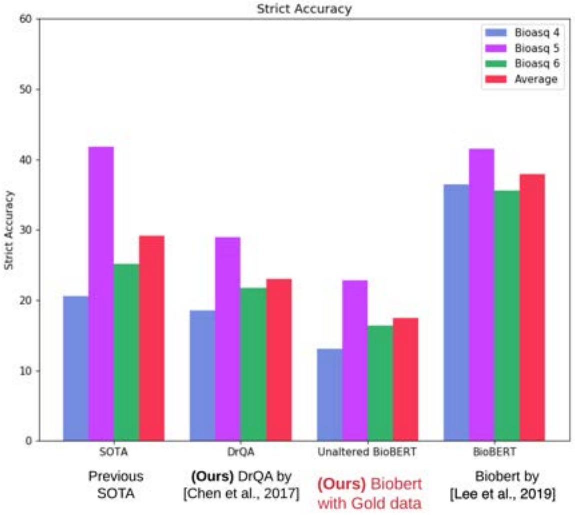


Figure: Modification of the paragraph text results in the variation of performance

Which dataset is better to Pre-train?

Some concerns

- How large is large enough for a dataset?
- Are synthetic datasets better than human annotated ones?
- What is the minimum size of a dataset required for optimal performance?
- How do we choose the best dataset for pre-training?

Some consequences

- BERT being the state of the art model (during the time of performing these experiments) is used in recent works.
- Large scale makes it harder to train the model from scratch with low resource compute, training and inference both take much longer times than previous neural models.

Pre-Training Datasets

Datasets	Train	Dev	Test
SQUAD v1.0	87,599	10,570	9,533
SQUAD v2.0	130,319	11,873	8,862
Hotpot QA	90,564	7,405	7,405
News QA	107,673	5,988	5,971

Table: Large scale datasets used in the experiments for pre-training, with their splits.

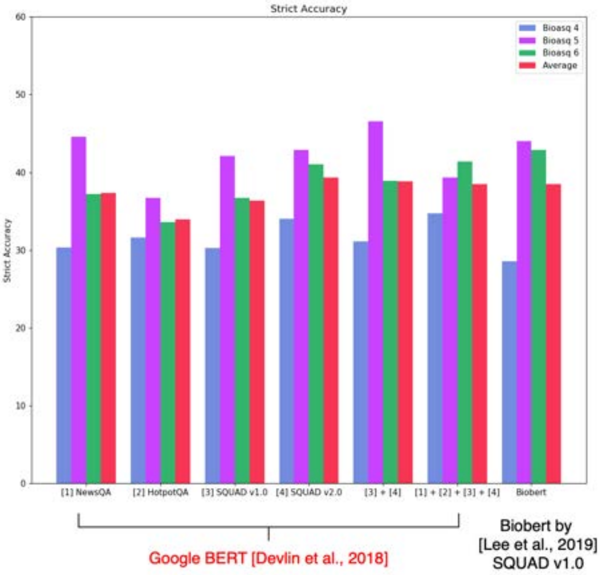


Figure: Experiments on different Pre-training datasets for Biomedical QA task.

Conclusion

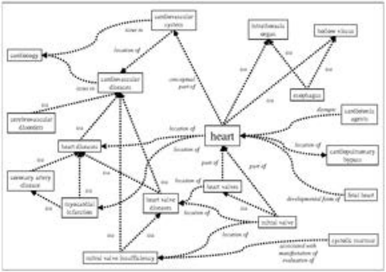
- Small scale datasets are suitable with deep learning models while using domain adaptation.
- Domain adaptation helps in improving QA performance.
- Open QA model performs better for modeling BIOASQ datasets.
- Different pre-training datasets have different impact on downstream domain adaptation performance.

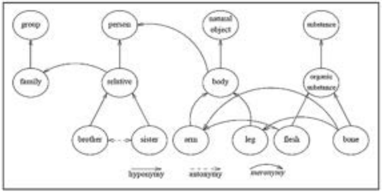
Plan

1. Introduction
2. State of the art
3. Building Domain-Specific Models
4. Leveraging Semantic Information
5. Conclusion

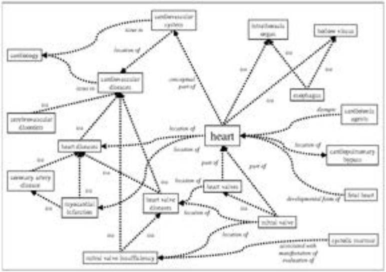
Semantic Information

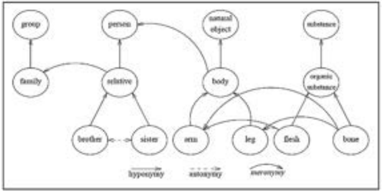
- End-to-end Neural models rely only on input and output to learn.
- Traditional QA pipeline methods rely on features from different sources such as named entities, part of speech tags, question types etc.
- "How can one build models which use best of the both worlds?"





Semantic and Structured Information





- Information about question types and expected answer types.
- Synonyms and variants of answers in the contexts.
- Information about entities from ontologies and knowledge bases.
- Syntactic features of words and sentences.

Approach 1 - Answer Variants

Q: Mutation of which gene is implicated in the Brain-lung-thyroid syndrome?

- Novel NKX2-1 Frameshift Mutations in Patients with Atypical Phenotypes of the Brain-Lung-Thyroid Syndrome.
- NKX2-1 mutations in brain-lung-thyroid syndrome: a case series of four patients.
- Brain-lung-thyroid syndrome (BLTS) characterized by congenital hypothyroidism, respiratory distress syndrome, and benign hereditary chorea is caused by thyroid transcription factor 1 (NKX2-1/TTF1) mutations.
- The disorder is caused by mutations to the NKX2.1 (TITF1) gene and also forms part of the \"brain-lung-thyroid syndrome\", in which additional developmental abnormalities of lung and thyroid tissue are observed.

A: thyroid transcription factor 1

Gold standard data annotated by experts - Misses many variants.

Using Answer Variants

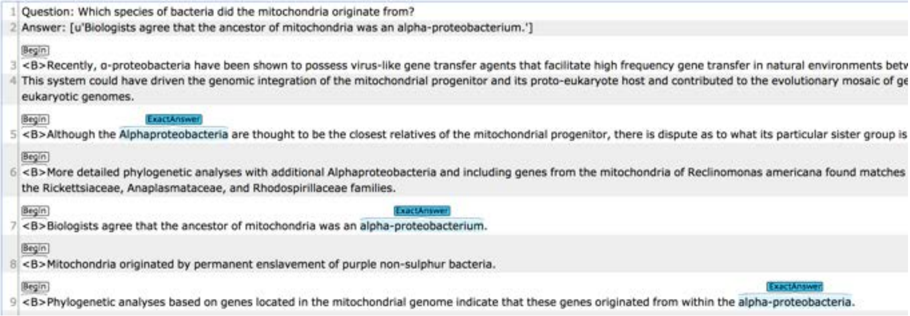
Q: Mutation of which gene is implicated in the Brain-lung-thyroid syndrome?

- Novel NKX2-1 Frameshift Mutations in Patients with Atypical Phenotypes of the Brain-Lung-Thyroid Syndrome.
- NKX2-1 mutations in brain-lung-thyroid syndrome: a case series of four patients.
- Brain-lung-thyroid syndrome (BLTS) characterized by congenital hypothyroidism, respiratory distress syndrome, and benign hereditary chorea is caused by thyroid transcription factor 1 (NKX2-1/TTF1) mutations.
- The disorder is caused by mutations to the NKX2.1 (TITF1) gene and also forms part of the \"brain-lung-thyroid syndrome\", in which additional developmental abnormalities of lung and thyroid tissue are observed.

A: thyroid transcription factor 1, NKX2-1, TTF1, TITF1

Several answer variants which are syntactically different but semantically represents the same entity are annotated by us.

Annotating Answer Variants



Manual Annotations

- Annotated using the Brat tool and UMLS meta thesaurus references. 618 questions were annotated manually.
- 3 people from CS background annotated for answer variants.
- Released this annotated dataset publicly `https://zenodo.org/record/1346193#.W3_WUZMzZQI`

Annotating Answer Variants

Gold standard answer : MDR - TB

Paragraph 1: Delamanid: a review of its use in patients with multidrug-resistant tuberculosis.

Paragraph 2: In conclusion, delamanid is a useful addition to the treatment options currently available for patients with MDR-TB.

Paragraph 3:
EXPERT OPINION: Delamanid showed potent activity against drug-susceptible and -resistant Mycobacterium tuberculosis in both in vitro and in vivo studies.

Automatic Annotations

- Annotated using the UMLS CUI identifiers from entities detected.
- CUI from gold standard answer and matching CUIs from paragraphs are mapped to find variants.

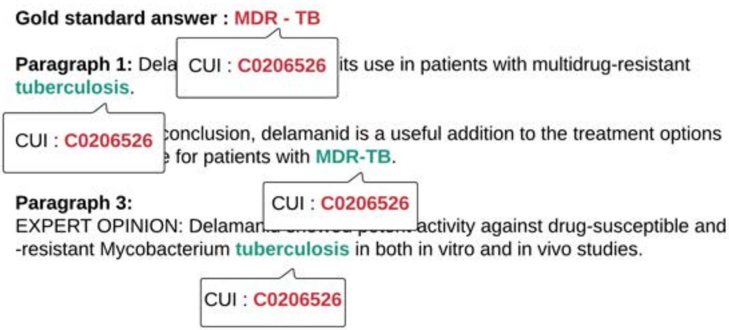
Annotating Answer Variants

```
Processing 00000000.tx.1: MDR-TB

Phrase: MDR-TB
>>>>> Phrase
mdr tb
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
  1000   C0206526:MDR-TB (Tuberculosis, Multidrug-Resistant) [Disease or Syndrome]
<<<<< Mappings
```

Automatic Annotations

- Annotated using the UMLS CUI identifiers from entities detected.
- CUI from gold standard answer and matching CUIs from paragraphs are mapped to find variants.

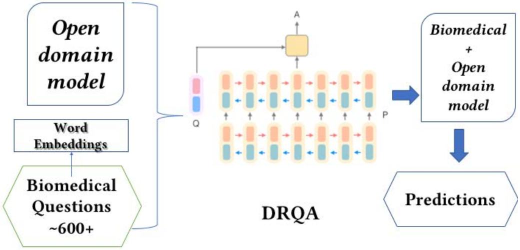
Annotating Answer Variants



Automatic Annotations

- Annotated using the UMLS CUI identifiers from entities detected.
- CUI from gold standard answer and matching CUIs from paragraphs are mapped to find variants.

Answer Variants and Performance Increase



Experiments

- Reading comprehension experiments with BIOASQ Gold Standard data and Annotated data.
- Comparison with Automatic and Manually annotated answers.
- SQUAD dataset for Open domain, BIOASQ dataset for biomedical domain.

Answer Variants and Performance Increase

Train set	5B					6B			
Finetune	Gold			Anno.		Gold		Anno.	
Eval	DeepQA	Gold	Anno.	Gold	Anno.	Gold	Anno.	Gold	Anno.
Strict	-	0.2551	**0.2962**	0.1666	**0.3333**	0.2669	**0.3090**	0.2265	**0.3948**
MRR	0.2620	0.3138	**0.3425**	0.2148	**0.4322**	0.3334	**0.3718**	0.2728	**0.4765**

Figure: 5-fold evaluation on 5B and 6B datasets of BIOASQ - [Kamath et al., 2018]

Evaluation	-	BIOASQ 5B		BIOASQ 6B	
-	Measures	Manual	Auto.	Manual	Auto.
Annotated Data	Strict	**45.83**	37.81	47.59	**48.59**
	MRR	**53.90**	46.55	58.39	**58.53**

Figure: Automatic vs Manually annotated answer datasets - (Our unpublished results)

Approach 2 - Expected Answer Types (EAT)

Biomedical Domain

Question: What disease in Loxapine prominently used for?
Answer: Schizophrenia
Expected Answer Type: Disease or Syndrome.
Semantic Group: DISO Disease or Syndrome.
Lexical Answer Type: disease

Question: Which drugs are utilized to treat amiodarone-induced thyroitoxicosis?
Answer: Antithyroid drugs
Expected Answer Type: Chemical Drug
Semantic Group: CHEM Chemicals & Drugs
Lexical Answer Type: drugs

Open Domain

Question: Which actor played the role of Walter White in the series Breaking Bad?
Answer: Bryan Cranston
Expected Answer Type (EAT): HUM - Person, Organisation
Lexical Answer Type (LAT): actor

Question: Where was the TV show Breaking Bad primarily filmed at?
Answer: Albuquerque
Expected Answer Type (EAT): LOC - Location
Lexical Answer Type (LAT): Where

Improving Answer Sentence Selection - Using EAT (Expected Answer Type)

-	Method	Question	Sentence
1	Original text	Who is the author of the book, 'The Iron Lady: a biography of Margaret Thatcher'	in 'The Iron Lady,' *Young* traces ...... the greatest woman political leader since *Catherine the Great*.
2	Replacement - (Tayyar Madabushi et al., 2018) (EAT Single type)	Who is the author of the book, 'The Iron Lady: a biography of Margaret Thatcher' *max_entity_left entity_left*	in 'The Iron Lady,' *max_entity_left* traces ...... the greatest woman political leader since *entity_left*.
3	EAT (Different types)	Who is the author of the book, 'The Iron Lady: a biography of Margaret Thatcher' *max_entity_left entity_hum*	in 'The Iron Lady,' *max_entity_left* traces ...... the greatest woman political leader since *entity_hum*.
4	EAT (MAX + Different types)	Who is the author of the book, 'The Iron Lady: a biography of Margaret Thatcher' *max_entity_hum entity_hum*	in 'The Iron Lady,' *max_entity_hum* traces ...... the greatest woman political leader since *entity_hum*.

Improving Answer Sentence Selection - Using EAT (Expected Answer Type)

Dataset	Split	#Plain Q	#EAT Q	#Entities
Trec QA	Train	1229	649 (52.8%)	13.96
	Dev	82	76 (92.68%)	5.02
	Test	100	82 (82%)	7.82

Table: Dataset annotated by [Madabushi et al., 2018] (EAT - Expected Answer Type)

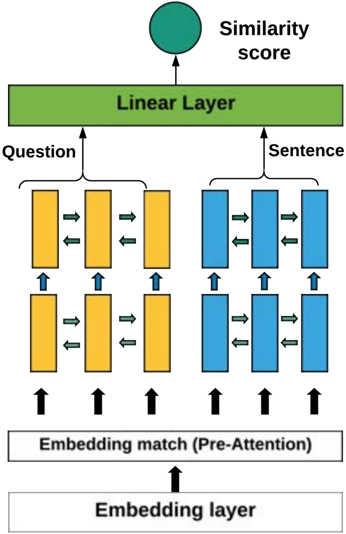


Figure: RNN-S model for Sentence Selection Task

Improving Answer Sentence Selection - Using EAT (Expected Answer Type)

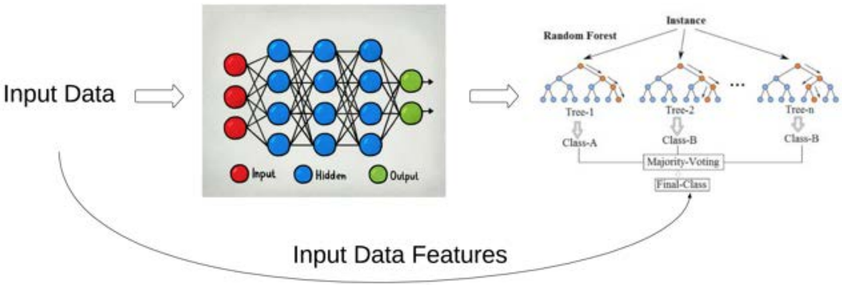
Datasets	Method	Acc.@1	MAP	MRR
TrecQA	Plain words - (Rao et al., 2016)	-	78	83.4
	EAT words - (Tayyar Madabushi et al., 2018)	-	83.6	86.2
	Plain words - RNN-S	78.95	80.24	84.81
	EAT words (single type) - RNN-S	85.26	85.28	**89.16**
	EAT words (different types) - RNN-S	85.26	**85.48**	88.11
	EAT words (MAX+different types) - RNN-S	**86.32**	85.42	88.86

Figure: Results using RNN-S model. EAT (Expected Answer Type) - [Kamath et al., 2019]

Approach 3 - Improving Top-1 Accuracy Using Semantic Features

```
"type": "factoid",
"body": "What is the function of the TMEM132 genes?",

"exact_answer":

        "TMEM132"

        "tandem immunoglobulin domains"

        "mutations associated with non - syndromic hearing loss , panic disorder and cancer ."

        "TMEM132 family , connecting the extracellular medium with the intracellular actin cytoskeleton"

        "tandem immunoglobulin domains"

        "TMEM132 family , connecting the extracellular medium with the intracellular actin cytoskeleton"
```

Improving Top-1 Accuracy Using Semantic Features



Method

- Neural network model outputs can be used to further improve QA performance using semantic and structured information.
- Top-K answers are re-ranked to obtain better Top-1 accuracy.
- Models such as *Multi Layer Perceptron, Adaboost, Random forests* were experimented using several features from OpenQA model [Lin et al., 2018], semantic and syntactic features and Expected Answer Type features.

Improving Top-1 Accuracy Using Semantic Features

Open domain QA Features

- Answer probability, Paragraph probability, Paragraph and answer length, Answer words overlap with paragraph words.
- Maximum value of answer probability, Maximum value of paragraph probability, Answer presence ratio, Summation of answer probability, Answer rank.
- Expected Answer Type (EAT) match, Cosine distance between Lexical Answer type word and answer words.

Biomedical QA additional Features

- Matching of Lexical Answer Type (LAT) word with UMLS Semantic Type match
- Matching of Lexical Answer Type (LAT) word with UMLS Semantic Group match
- Matching of Lexical Answer Type (LAT) word with UMLS CUI match

Improving Top-1 Accuracy Using Semantic Features

Algorithm	Accuracy
PSPR - Baseline	41.1
Randomforest	42.86
Adaboost	**43.23**
MLP classifier	42.23

Features removed from the best model	Accuracy	Drop
Best model score	43.23	-
- Maximum value of paragraph probability + Answer presence ratio	41.36	**1.87**
- Maximum value of paragraph probability	42.23	**1**
- Summation of Ans Prob. + Answer presence ratio	42.46	**0.77**
- Answer presence ratio	42.53	**0.7**
- Summation of answer probability	43.03	**0.2**
- Answer Probability	43.06	**0.17**
- Sentence Probability	43.2	**0.03**

Figure: Results on the QUASAR-T dataset (Open Domain)

Improving Top-1 Accuracy Using Semantic Features

Datasets	Finetune	BIOASQ 4	BIOASQ 5	BIOASQ 6
Baseline	Strict	30.31	46.83	42.79
	Lenient	45.00	52.66	53.41
Adaboost	Strict	**39.37**	44.00	**45.96**
	Lenient	45.00	52.66	53.41
Randomforest	Strict	**38.75**	46.00	**46.58**
	Lenient	45.00	52.66	53.41
MLP	Strict	**34.37**	46.00	38.50
	Lenient	45.00	52.66	53.41

Features removed from the best model	Bioasq 5 Acc.	Bioasq 6 Acc.
Best model score	46.00	46.58
- Maximum value of answer probability + Answer overlap	42.66 (**3.34**)	44.72 (**1.86**)
- Maximum value of answer probability	43.33 (**2.67**)	44.72 (**1.86**)
- Paragraph Probability	44.66 (**1.34**)	45.96 (**0.62**)
- LAT Semantic Type	45.33 (**0.67**)	45.96 (**0.62**)
- Answer overlap	45.33 (**0.67**)	45.96 (**0.62**)

Figure: Results on the BIOASQ dataset (Biomedical Domain)

Conclusion

- Annotated answer variants show improvement in performance for Reading Comprehension.
- Explicitly highlighting Expected Answer Types (EAT) in the data helps in improving certain QA tasks.
- Top-1 accuracy can be improved using Semantic information.


That's all Folks!

Plan

1 Introduction

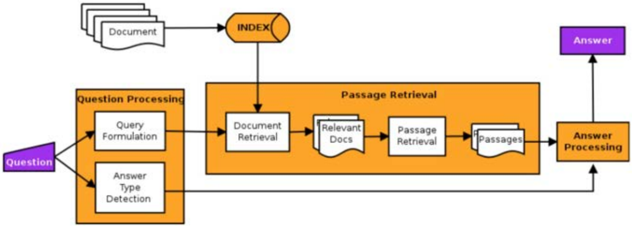
2 State of the art

3 Building Domain-Specific Models

4 Leveraging Semantic Information

5 Conclusion

Whole QA pipeline using different QA models



QA Pipeline using SOTA models

- Can the current State of the Art models replace the pipeline with an end-to-end model?
- How does the performance compare with neural models which use pipeline architecture?

Whole QA pipeline using SOTA models

-	Model	Accuracy
1	SQUAD using DRQA (Chen et al., 2017)	69.5
2	Open QA (BM25) and BERT model (K. Lee et al., 2019)	28.1
3	Open QA using LSTM model (Chen et al., 2017)	27.1
4	Whole QA using BERT model (K. Lee et al., 2019)	26.5

QA Pipeline using SOTA models

- BERT end-to-end model for whole QA pipeline performs **worse** compared to LSTM model DRQA which is end-to-end.
- BM25 for document retrieval and BERT for QA performs better than BERT alone.

Whole QA Pipeline on TrecQA data

Model	Accuracy
Tree kernels methods by (Severyn and Moschitti, 2013)	70.8
Sequence tagging model by (Yao et al., 2013)	67.2
Reading Comprehension mode	58.02
OpenQA pre-trained with QUASAR-T	49.38
OpenQA on TrecQA data only	45.68

QA Pipeline on TrecQA data

- Using the OpenQA model by [Lin et al., 2018] on TrecQA task dataset for Answer Extraction.
- Non-neural models performed better.

Large Scale Language Models - Concerns

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Some concerns

- Expensive compute required. BERT took 4 days on 16 cloud TPUv3 (64 chips) which costs around $14,000.
- CO2 emissions are a side effect which are often neglected into consideration. [Strubell et al., 2019]

Large Scale Language Models - Concerns

Conversation B

Human:	*Hi!*
Meena:	*Hi! How are you this fine evening?*
Human:	*I'm doing well. it's morning my time!*
Meena:	*Good morning! What are your plans for today?*
Human:	*Get my work done then watch a movie at home in the night*
Meena:	*Sounds fun! What movie?*
Human:	*I haven't decided yet, perhaps an Oscar-nominated one. any suggestion?*
Meena:	*I'm not too familiar with recent Oscar winners, but I do know I wanted to see The Grand Budapest Hotel.*
Human:	*ah, tell me about that movie!*
Meena:	*It's from the same director and writer as Moonrise Kingdom. Wes Anderson is pretty well known for making funny, colorful films. I've heard it's pretty good.*
Human:	*ha! I've searched on the internet and both movies are very colorful!*
Meena:	*I'm looking forward to it. I love colorful movies.*

Some concerns

- A new chatbot model by Google AI [Adiwardana et al., 2020] costs $1,400,000 to train for 30 days on the TPU cloud.
- Google uses renewable energy sources for cloud, what about others?

Conclusion

- Domain Adaptation
  - → From Open domain towards Biomedical domain QA using several techniques.
- Semantic information
  - → Explicit use of semantic and structured information can help.
- Choose simple models most of the times.
  - → Simple models are better for several reasons.
- Contributions
  - → Several SOTA QA models built, modified and experimented.
  - → Annotated datasets and codes released publicly.

Future Perspectives

- Large Scale Language Models - a new milestone in NLP.
  $\rightarrow$ Building smaller models which perform similarly.
- Low resource domain data adaptation.
  $\rightarrow$ Expert domains might not always have large scale text for pre-training.
- Fusion of structured and semantic information from ontologies and knowledge bases into language models.


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
BİYAN
SHUKRIA
TINGKI
YAQHANYELAY
TASHAKKUR ATU
SUKSAMA
EKHMET
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
KOMAPSUMNIDA
MAAKE
GRAZIE
MEHRBANI
PALDIES
BOLZİN
MERCI

Publications

- **2019 (Co-authored publication) - Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations** - Anna Koroleva, Sanjay Kamath, Patrick Paroubek. Journal of Biomedical Informatics: X, Published in October 2019.
- **2019 - How to Pre-Train Your Model? Comparison of Different Pre-Training Models for Biomedical Question Answering.** - Proceedings of the 7th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering. ECMLPKDD, September 2019.
- **2019 - Predicting and Integrating Expected Answer Types into a Simple Recurrent Neural Network Model for Answer Sentence Selection.** - 20th International Conference on Computational Linguistics and Intelligent Text Processing - CICLING 2019, April 2019.
- **2018 - An Adaption of BIOASQ Question Answering dataset for Machine Reading systems by Manual Annotations of Answer Spans.** - Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering. EMNLP, October 2018.
- **2018 - Verification of the Expected Answer Type for Biomedical Question Answering.** - HQA workshop, companion proceedings of the The Web Conference 2018, April 2018.
- **2017 - A Study of Word Embeddings for Biomedical Question Answering.** - 4e édition du Symposium sur l'Ingénierie de l'Information Médicale, November 2017.