Probabilistic Neural Topic Models for Text Understanding

Gabriele Pergola
Department of Computer Science
University of Warwick

16th April, 2021
Motivation
Research Objectives
Contributions
Publications
Discussion
Future Work and Conclusion
Motivation

Making sense of text in **large corpora**

- User Reviews, Electronic Health Records, Social Networks, etc.
Motivation

Making sense of text in **large corpora**
- User Reviews, Electronic Health Records, Social Networks, etc.

Dealing with **linguistic features** across different domains and contexts
- User reviews might be characterized by colloquial *idioms, slang or contractions*
Motivation

Making sense of text in **large corpora**
- User Reviews, Electronic Health Records, Social Networks, etc.

Dealing with **linguistic features** across different domains and contexts
- User reviews might be characterized by colloquial **idioms**, **slang** or **contractions**

- **Stars**: Our children didn’t manage to **clean their plates!** Plenty of food!

- **Stars**: After one cycle the crockery is still dirty, it doesn’t **clean the plates** even at full power.
Motivation

Making sense of text in **large corpora**

- User Reviews, Electronic Health Records, Social Networks, etc.

Dealing with **linguistic features** across different domains and contexts

- User reviews might be characterized by colloquial *idioms, slang* or *contractions*
- Clinical notes often contain *technical jargon, multiword phrases*, medical
  *abbreviations* and *polysemous* terms

*White blood cell*

*Shortness of breath*

…
Motivation

Making sense of text in **large corpora**
- User Reviews, Electronic Health Records, Social Networks, etc.

Dealing with **linguistic features** across different domains and contexts
- User reviews might be characterized by colloquial *idioms, slang* or *contractions*
- Clinical notes often contain *technical jargon*, *multiword phrases*, medical *abbreviations* and *polysemous* terms
- Authors mix their *opinions* with *factual descriptions*

⭐⭐⭐⭐⭐

Overall, this is a good film and an excellent adaption. [...] the myriad inhabitants of Middle-earth, the legendary Rings of Power, and the fellowship of hobbits, elves, dwarfs, and humans—led by the wizard Gandalf (Ian McKellen) and the brave hobbit Frodo.
Motivation

- **Topic models** have established themselves as effective tools to generate concise and expressive representations of high volumes of documents.
  - *Global Context* - Document-based embedding (e.g. LDA, LSI).

"An introduction to Topic models", M. J. Paul 2013
Motivation

- **Topic models** have established themselves as effective tools to generate concise and expressive representations of high volumes of documents.
  - *Global Context* - Document-based embedding (e.g. LDA, LSI).

- Distributional representations of **word** and **language models**
  - *Local Context* - Window-based embedding (e.g. word2vec, BERT).

Word Embeddings

...government debt problems turning into banking crises as happened in 2009... saying that Europe needs unified banking regulation to replace the hodgepodge... India has just given its banking system a shot in the arm...

These context words will represent *banking*

Masked Language Models

the man went to the [MASK] to buy a [MASK] of wine
Our hypothesis is that **topic models** and **neural architectures** are a suitable combination for capturing text semantic, with benefits connected to several NLP open problems.

**RO 1**  Combining global and local context of words
Our hypothesis is that topic models and neural architectures are a suitable combination for capturing text semantic, with benefits connected to several NLP open problems.

**RO 1** Combining global and local context of words
Our hypothesis is that **topic models** and **neural architectures** are a suitable combination for capturing text semantic, with benefits connected to several NLP open problems.

**RO 1**  Combining global and local context of words

**RO 2**  Generating fine-grained topics.
Our hypothesis is that **topic models** and **neural architectures** are a suitable combination for capturing text semantic, with benefits connected to several NLP open problems.

**RO 1** Combining global and local context of words

**RO 2** Generating fine-grained topics.

**RO 3** Incorporating unstructured knowledge.
Our hypothesis is that **topic models** and **neural architectures** are a suitable combination for capturing text semantic, with benefits connected to several NLP open problems.

**RO 1** Combining global and local context of words

**RO 2** Generating fine-grained topics.

**RO 3** Incorporating unstructured knowledge.

**RO 4** Incorporating structured knowledge.

**Domain-specific lexicons and concepts**

Patients ** ENTITY with diabetes ** ENTITY (HR ** ENTITY 1.59) were more likely to reach to the ** composite ENTITY endpoints ** ENTITY than those without.
Our hypothesis is that topic models and neural architectures are a suitable combination for capturing text semantic, with benefits connected to several NLP open problems.

**Research Objectives**

RO 1  Combining global and local context of words

RO 2  Generating fine-grained topics.  

RO 3  Incorporating unstructured knowledge.

RO 4  Incorporating structured knowledge.

RO 5  Evaluation on downstream tasks.

<table>
<thead>
<tr>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment Analysis</td>
</tr>
<tr>
<td>Aspect Extraction</td>
</tr>
<tr>
<td>Biomedical QA</td>
</tr>
</tbody>
</table>
Contributions

“TDAM: a Topic-Dependent Attention Model for Sentiment Analysis”
*Information Processing and Management, 2019*

Research Objectives

RO 1  Combining global and local context of words

RO 2  Generating fine-grained topics.

RO 3  Incorporating unstructured knowledge.

RO 4  Incorporating structured knowledge.

RO 5  Evaluation on downstream tasks.

Contributions

*External memory* keeping track of word co-occurrences

*Aspect* extraction through sentiment topics

*Word embeddings* induced from sentiment polarity

*Sentiment classification and Aspect Extraction*
Contributions

“A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews”

North American Chapter of the Association for Computational Linguistics (NAACL), 2021

Research Objectives

RO 1 Combining global and local context of words

RO 2 Generating fine-grained topics.

RO 3 Incorporating unstructured knowledge.

RO 4 Incorporating structured knowledge.

RO 5 Evaluation on downstream tasks.

Contributions

Initializing with pre-training word embedding

Generation of disentangled topics

MOBO dataset and introduction of disentangling rate for topics
## Contributions

**“Topical Phrase Extraction from Clinical Reports by Incorporating both Local and Global Context”**

*The 2nd AAAI Workshop on Health Intelligence, AAAI 2018*

<table>
<thead>
<tr>
<th>Research Objectives</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RO 1    Combining global and local context of words</td>
<td>Modified <em>inference process</em> joining word embedding information</td>
</tr>
<tr>
<td>RO 2    Generating fine-grained topics.</td>
<td>Identification of <em>topical phrases</em> encoded by character-based word embedding</td>
</tr>
<tr>
<td>RO 3    Incorporating unstructured knowledge.</td>
<td><em>Domain-specific</em> word embedding</td>
</tr>
<tr>
<td>RO 4    Incorporating structured knowledge.</td>
<td>Identification of <em>topical phrases</em> encoded by character-based word embedding</td>
</tr>
</tbody>
</table>
Contributions

“Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies”
European Chapter of the Association for Computational Linguistics (EACL) 2021

Research Objectives

RO 1 Combining global and local context of words

RO 2 Generating fine-grained topics.

RO 3 Incorporating unstructured knowledge.

RO 4 Incorporating structured knowledge.

RO 5 Evaluation on downstream tasks.

Contributions

Biomedical entity detection via a scientific NER model

Impact on several language models
Publications

“A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews”
North American Chapter of the Association for Computational Linguistics (NAACL), 2021

“Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies”
European Chapter of the Association for Computational Linguistics (EACL), 2021

“TDAM: a Topic-Dependent Attention Model for Sentiment Analysis”
Information Processing and Management, 2019

“Topical Phrase Extraction from Clinical Reports by Incorporating both Local and Global Context”
The 2nd AAAI Workshop on Health Intelligence (AAAI), 2018
“Adversarial Learning of Poisson Factorisation Model for Gauging Brand Sentiment in User Reviews”
Runcong Zhao, Lin Gui, Gabriele Pergola, Yulan He.
European Chapter of the Association for Computational Linguistics (EACL), 2021

“CHIME: Crosspassage Hierarchical Memory Network for Generative Review Question Answering”
Junru Lu, Gabriele Pergola, Lin Gui, Binyang Li, Yulan He.
The 28th International Conference on Computational Linguistics (COLING), 2020

“Neural Topic Model with Reinforcement Learning”
Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, Yulan He.
The 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019
TDAM: a Topic-Dependent Attention Model for Sentiment Analysis

Gabriele Pergola, Lin Gui, Yulan He
University of Warwick

Information Processing and Management
2019

Chosen for poster presentation at EurNLP
Outline

• What is Sentiment Analysis? What are “Topics”?

• Problem - How to Extract Polarity-Bearing Topics from Text using Neural Models

• The Consciousness Prior

• TDAM: A Topic-Dependent Attention Model

• Summary of Results
Sentiment Analysis

Aspect-Based Sentiment Analysis

- Nice view of river.
- But sushi rolls are bad.
Topic Modeling - 1

sick  
sore  
throat
feel
fever
flu ...  

allergies  
nose
eyes
allergy
allergic
sneezing ...

watch  
watching
tv
killing
movie
seen ...

class
school
read
test
doing
finish ...

Michael Paul @mjp39 · Jan 24
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv

"An introduction to Topic models", M. J. Paul 2013
Topic Modeling - 2

Michael Paul @mjp39 · Jan 24
I've had the flu and fever all week :( staying home from school and watching a lot of tv
Word representations should depend on the **topical context** in which they appear in. This context emerges from the co-occurrence of words across documents.

Our children didn’t manage to **clean their plates**! Plenty of food!

After one cycle the crockery is still dirty, it doesn’t **clean the plates** even at full power.
We want to extract topics by looking at the attention weights learned by neural models; however, there is no mechanism to separate words into multiple clusters representing polarity-bearing topics.
We want to extract **topics** by looking at the **attention weights** learned by neural models; however, there is no mechanism to separate words into multiple clusters representing polarity-bearing topics.
The Consciousness Prior

<< [...] the consciousness RNN $C()$ a tool for exploring interpretations or plans or to sample predictions about the future. We can also think of the consciousness RNN as the tool to isolate a particular high-level abstraction and extract the information about it >> [1].

TopicGRU – Attention to extract topics

**Representation State**

$$h_t = F(s_t, h_{t-1})$$

**Conscious State**

$$c_t = C(h_t, c_{t-1}, z_t)$$

**Modified GRU**

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + V_r q_{t-1})$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + V_z q_{t-1})$$

$$\hat{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1} + V_h q_{t-1}))$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t$$

$$u_{it} = \tanh(W_w h_{it})$$

$$\alpha_{it}^k = \text{softmax}(u_{it}^T e_k)$$

$$q_{it} = \sum_{k=1}^{K} \alpha_{it}^k \otimes e_k$$
Topic Dependent Attention Model
Topic Dependent Attention Model
Topic Dependent Attention Model

FROM SENTENCES TO DOCUMENTS

FROM WORDS TO SENTENCES
Topic Dependent Attention Model

FROM SENTENCES TO DOCUMENTS

FROM WORDS TO SENTENCES
# Sentiment Classification

<table>
<thead>
<tr>
<th>Methods</th>
<th>Yelp 18</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM</td>
<td>74.5 ± 0.2</td>
<td>72.1 ± 0.2</td>
</tr>
<tr>
<td>BiLSTM - Mtl</td>
<td>74.2 ± 0.2</td>
<td>71.8 ± 0.1</td>
</tr>
<tr>
<td>BiGRU</td>
<td>75.5 ± 0.1</td>
<td>72.5 ± 0.3</td>
</tr>
<tr>
<td>BiGRU - Mtl</td>
<td>75.4 ± 0.2</td>
<td>72.1 ± 0.3</td>
</tr>
<tr>
<td>HAN</td>
<td>83.7 ± 0.2</td>
<td>78.4 ± 0.2</td>
</tr>
<tr>
<td>HAN - Mtl</td>
<td>83.6 ± 0.3</td>
<td>78.2 ± 0.3</td>
</tr>
<tr>
<td>S-LDA</td>
<td>70.8 ± 0.2</td>
<td>64.6 ± 0.1</td>
</tr>
<tr>
<td>SCHOLAR</td>
<td>77.3 ± 0.2</td>
<td>71.4 ± 0.2</td>
</tr>
<tr>
<td>TDAM</td>
<td>84.2 ± 0.2</td>
<td>78.9 ± 0.2</td>
</tr>
<tr>
<td>TDAM - Mtl</td>
<td><strong>84.5 ± 0.3</strong></td>
<td><strong>79.1 ± 0.2</strong></td>
</tr>
</tbody>
</table>
## Sentiment Classification

<table>
<thead>
<tr>
<th>Methods</th>
<th>Yelp 18</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM</td>
<td>74.5 ± 0.2</td>
<td>72.1 ± 0.2</td>
</tr>
<tr>
<td>BiLSTM - Mtl</td>
<td>74.2 ± 0.2</td>
<td>71.8 ± 0.1</td>
</tr>
<tr>
<td>BiGRU</td>
<td>75.5 ± 0.1</td>
<td>72.5 ± 0.3</td>
</tr>
<tr>
<td>BiGRU - Mtl</td>
<td>75.4 ± 0.2</td>
<td>72.1 ± 0.3</td>
</tr>
<tr>
<td>HAN</td>
<td>83.7 ± 0.2</td>
<td>78.4 ± 0.2</td>
</tr>
<tr>
<td>HAN - Mtl</td>
<td>83.6 ± 0.3</td>
<td>78.2 ± 0.3</td>
</tr>
<tr>
<td>S-LDA</td>
<td>70.8 ± 0.2</td>
<td>64.6 ± 0.1</td>
</tr>
<tr>
<td>SCHOLAR</td>
<td>77.3 ± 0.2</td>
<td>71.4 ± 0.2</td>
</tr>
<tr>
<td>TDAM</td>
<td>84.2 ± 0.2</td>
<td>78.9 ± 0.2</td>
</tr>
<tr>
<td>TDAM - Mtl</td>
<td>84.5 ± 0.3</td>
<td>79.1 ± 0.2</td>
</tr>
</tbody>
</table>
Aspect-based Sentiment Analysis

Detecting aspect (e.g. food, service, options) and polarity (positive, negative, neutral)
## Aspect-based Sentiment Analysis

Detecting **aspect** (e.g. food, service, options) and **polarity** (positive, negative, neutral)

### Positive polarity - Food/#Quality

<table>
<thead>
<tr>
<th>Text</th>
<th>Aspect</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) the food was all good but it was way too</td>
<td>FOOD#QUALITY</td>
<td>pos</td>
</tr>
<tr>
<td>2) the pizza’s are light and scrumptious</td>
<td>RESTAURANT#MISCELLANEOUS</td>
<td>neg</td>
</tr>
<tr>
<td>3) the food is great and they make a mean bloody mary</td>
<td>FOOD#STYLE.OPTIONS</td>
<td>neg</td>
</tr>
<tr>
<td>4) great draft and bottle selection and the pizza rocks</td>
<td>RESTAURANT#MISCELLANEOUS</td>
<td>neut</td>
</tr>
<tr>
<td>5) the food is simply unforgettable!</td>
<td>FOOD#QUALITY</td>
<td>pos</td>
</tr>
</tbody>
</table>

### Negative polarity - Food/#Quality

<table>
<thead>
<tr>
<th>Text</th>
<th>Aspect</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) the pancakes were certainly inventive but $8.50</td>
<td>FOOD#STYLE.OPTIONS</td>
<td>neg</td>
</tr>
<tr>
<td>for 3 - 6&quot; pancakes (one of them was more like 5&quot;)</td>
<td>RESTAURANT#MISCELLANEOUS</td>
<td>neg</td>
</tr>
<tr>
<td>2) a beautiful assortment of enormous white gulf prawns,</td>
<td>FOOD#QUALITY</td>
<td>neg</td>
</tr>
<tr>
<td>smoked albacore tuna, [...] and a tiny pile of dungeness</td>
<td>RESTAURANT#MISCELLANEOUS</td>
<td>neg</td>
</tr>
<tr>
<td>3) space was limited, but the food made up for it</td>
<td>FOOD#STYLE.OPTIONS</td>
<td>neutr</td>
</tr>
<tr>
<td>4) the portions are big though, do not order too much</td>
<td>FOOD#QUALITY</td>
<td>pos</td>
</tr>
<tr>
<td>5) not the biggest portions but adequate.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Aspect-based Sentiment Analysis

Detecting aspect (e.g. food, service, options) and polarity (positive, negative, neutral)
Conclusion

• A possible extension of the hierarchical model is to introduce an intermediate level between words and sentences, defining a *discourse-level layer*. This discourse-level layer would determine the **elementary discourse units (EDUs)**.

• Use a **Neural Topic Model** to jointly update the external memory inferring topical representations.

• Employment of **Contextualized Language Models** in place of GloVe embedding.
Disentangled Adversarial Topic Model

Gabriele Pergola, Lin Gui, Yulan He
University of Warwick

“A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews”
North American Chapter of the Association for Computational Linguistics (NAACL)
2021
Problem: **Disentangling Topics**

- The application of topic models to text with users’ opinions (e.g. book reviews) often leads to the generation of **topics mixing different aspects**. For example, information about book’s plot or characters gets mixed with user’s feelings and opinions.

- Therefore, we want to separate **polarity-bearing topics** from **neutral topics**.
  
  Providing features disentangling different aspects in text.

- As a case study, we applied our model to movie and book reviews to disentangle topics about their **plots** and topics about the users’ **opinions**.

  Fig. 1 shows an example of **polarity-disentangled topics**.

---

**Fig. 1**: Disentangled topics extracted by DIATOM from the Amazon reviews for “The Hobbit”.
We propose to combine a **Neural Topic Model** architecture with **adversarial training** to disentangle polarity-bearing topics from neutral ones.
Model: **Neural Topic Model with Adversarial Training**

We propose to combine a **Neural Topic Model** architecture with **adversarial training** to disentangle polarity-bearing topics from neutral ones.

- The **VAE architecture** extracts topics through a bottleneck mechanism, reading and reconstructing the documents.
Model: **Neural Topic Model with Adversarial Training**

We propose to combine a **Neural Topic Model** architecture with **adversarial training** to disentangle polarity-bearing topics from neutral ones.

- The **VAE architecture** extracts topics through a bottleneck mechanism, reading and reconstructing the documents.

- An **adversarial mechanism** separate polarised and neutral features, corresponding to polarised and neutral topics.
Model: Neural Topic Model with Adversarial Training

We propose to combine a Neural Topic Model architecture with adversarial training to disentangle polarity-bearing topics from neutral ones.

- The VAE architecture extracts topics through a bottleneck mechanism, reading and reconstructing the documents.

- An adversarial mechanism separate polarised and neutral features, corresponding to polarised and neutral topics.

- An analogous mechanism induces a prevalence of plot topics among all the possible neutral aspects.
We also introduce a new dataset, namely the **MOBO dataset**, made up of MOvie and BOok reviews, paired with their related plots. The reviews come from different publicly available datasets: IMDB, GoodReads and Amazon reviews.

Additionally, around **15,000 sentences** of these reviews were manually annotated and used (only) during the evaluation phase to compute the disentanglement rate between polarity-bearing topics and plot (or neutral) topics.

```
<table>
<thead>
<tr>
<th>Statistics</th>
<th>IMDB</th>
<th>GoodReads</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of plots</td>
<td>1,131</td>
<td>150</td>
<td>100</td>
</tr>
<tr>
<td>No. of reviews</td>
<td>25,836</td>
<td>83,852</td>
<td>32,375</td>
</tr>
<tr>
<td>No. of reviews / plot (avg / max / min)</td>
<td>24 / 30 / 10</td>
<td>954 / 3,000 / 549</td>
<td>464 / 1525 / 272</td>
</tr>
<tr>
<td>Pos / Neg / Neutral distribution</td>
<td>0.46 / 0.54 / 0</td>
<td>0.33 / 0.50 / 0.17</td>
<td>0.32 / 0.46 / 0.22</td>
</tr>
<tr>
<td>Training set</td>
<td>20,317</td>
<td>65,816</td>
<td>25,883</td>
</tr>
<tr>
<td>Development set</td>
<td>2,965</td>
<td>9,007</td>
<td>3,275</td>
</tr>
<tr>
<td>Test set</td>
<td>2,554</td>
<td>9,029</td>
<td>3,217</td>
</tr>
<tr>
<td>No. of annotated sentences</td>
<td>6,000</td>
<td>6,000</td>
<td>6,000</td>
</tr>
</tbody>
</table>
```
To automatically evaluate whether a topic is polarised or neutral, we perform **topic labelling** using the annotated sentences in the MOBO dataset.

For each topic we compute a **topic embedding** taking the normalized weighted average of the word embeddings, and we then retrieve the top 10 most similar human-annotated sentences via cosine similarity between the topic embedding and each **sentence embedding**, where the sentence embedding is computed using the **Sentence-BERT** encoder.

The most frequent label among the retrieved sentences is adopted as the topic’s label.
Results: **Topic Evaluation - 2**

- **Gandalf**
- **Frodo**
- **Jackson**
- **Tolkien**
- **Dwarf**
- **Hobbit**

**0.619 Plot**

the myriad inhabitants of Middle-earth, the legendary Rings of Power, and the fellowship of **hobbits**, elves, **dwarfs**, and humans–led by the wizard **Gandalf** (Ian McKellen) and the brave **hobbit Frodo**.

**0.523 Plot**

This is the beginning of a trilogy; soon to be finalized.

**0.597 Positive**

Overall, this is a good film and an excellent adaption.

**0.495 Positive**

It’s great acting, superb cinematography and excellent writing.

*Fig. 3: Examples of topics with the assigned sentences. Each sentence has a manually assigned label (i.e. Positive, Negative, Plot or None)*
We analyse how the proportion of polarity-bearing topics varies across standard and sentiment topic models.

We notice that despite the signal from the document labels, sLDA and SCHOLAR tend to produce topics rather balanced in terms of neutral and polarity-bearing topics. Joint-Sentiment Topic model has a more skewed distribution towards opinion topics.

DIATOM instead generates a proportion of opinion topics approaching the expected proportion set up by the model, demonstrating the capability to control the generation of neutral and polarity-bearing topics.

Results: Disentangling rate
In our current model, the latent plot topics extracted from reviews are encouraged to have a similar discriminative power as the latent topic learned from plots directly for predicting the plots. It is also possible to impose a Gaussian prior centred on “z_d” for the latent plot topics in reviews instead of using the Gaussian prior of zero mean and unit variance.

Another approach would consist of replacing the plot classifier with a discriminator as typically used in GAN training that the learned plot topics from different sources (reviews and plots) are competing.

An additional adversarial mechanism could be employed to differentiate opinion topics based on their overall polarity, to avoid topics with mixed sentiments.
Making sense of clinical notes by exploiting local and global context of words

Gabriele Pergola, Yulan He, David Lowe

The 2nd AAAI Workshop on Health Intelligence (AAAI)

2018
Outline

• Making Sense of Text (focusing on clinical documents)
• Topic Modeling and Distributed Representations of Language
• Context-GPU: Combining Local and Global Context
• Conclusions and Future Works
Natural language processing for medical documents

Find me relevant documents on this topic!

Find me relevant facts about this issue!

Find me scientific papers which help treat this patient!

I need more information on my health problem
Research Challenges

Is my inflammation your inflammation?

[1]: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Research Challenges

- **COMPOSITIONAL SEMANTIC** ‘spike protein sequence’ and ‘white blood cell’.
Research Challenges

- **COMPOSITIONAL SEMANTIC**
  
  ‘spike protein sequence’ and ‘white blood cell’.

- **TECHNICAL JARGON**
  
  ‘basophilia’, ‘synechococcus elongatus’.

Research Challenges

- **Compositional Semantic**: ‘spike protein sequence’ and ‘white blood cell’.
- **Technical Jargon**: ‘basophilia’, ‘synechococcus elongatus’.
- **Abbreviations**: S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’ - ‘systolic blood pressure’
Research Challenges

- **Compositional Semantic**
  - ‘spike protein sequence’ and ‘white blood cell’.

- **Technical Jargon**
  - ‘basophilia’, ‘synechococcus elongatus’.

- **Abbreviations**
  - S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’
    - ‘systolic blood pressure’

- **Polysemy**
  - ‘Inflammation’ up to 5 different meanings\(^1\).

[1]: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Research Challenges

• **COMPOSITIONAL SEMANTIC**
  
  ‘spike protein sequence’ and ‘white blood cell’.

• **TECHNICAL JARGON**
  
  ‘basophilia’, ‘synechococcus elongatus’.

• **ABBREVIATIONS**
  
  S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’
  - ‘systolic blood pressure’

• **POLYSEMY**
  
  ‘Inflammation’ up to 5 different meanings\(^1\).

• **LACK OF STRUCTURE**

---

[1]: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Research Challenges

• **COMPOSITIONAL SEMANTIC**
  
  ‘spike protein sequence’ and ‘white blood cell’.

• **TECHNICAL JARGON**
  
  ‘basophilia’, ‘synechococcus elongatus’.

• **ABBREVIATIONS**
  
  S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’
  
  - ‘systolic blood pressure’

• **POLYSEMY**
  
  ‘Inflammation’ up to 5 different meanings¹.

• **LACK OF STRUCTURE**

• **DATA AVAILABILITY**

---

¹: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Research Challenges

- **Compositional Semantic**
- **Technical Jargon**
- **Abbreviations**
- **Polysemy**
- **Lack of Structure**
- **Data Availability**

**Topical phrases**
Topics made of relevant phrases

**Contexts**
Word local and global context

**External source of knowledge**
Embedding exploited within the inference process

[1]: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Topic Modeling - 1

- sick
- sore
- throat
- feel
- fever
- flu
- ... 
- allergies
- nose
- eyes
- allergy
- allergic
- sneezing
- ... 
- watch
- watching
- tv
- killing
- movie
- seen
- ... 
- class
- school
- read
- test
- doing
- finish
- ... 

Michael Paul @mjp39 · Jan 24
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv

Credits for this slide to “An introduction to Topic models”, M. J. Paul 2013
Topic Modeling - 2

Michael Paul @mjp39 · Jan 24
I’ve had the flu and fever all week :( staying home from school and watching a lot of tv
Word Embedding

Distributional hypothesis: “You shall know a word by the company it keeps”
J. R. Firth, British Linguist, 1957

I love NLP and I like dogs

\[
I = [0 \ 1 \ 0 \ 1 \ 1 \ 0] \\
Love = [1 \ 0 \ 1 \ 0 \ 0 \ 0] \\
NLP = [0 \ 1 \ 0 \ 1 \ 0 \ 0] \\
And = [1 \ 0 \ 1 \ 0 \ 0 \ 0] \\
Like = [1 \ 0 \ 0 \ 0 \ 0 \ 1] \\
Dogs = [0 \ 0 \ 0 \ 0 \ 1 \ 0]
\]
Word Embedding

Distributional hypothesis: “You shall know a word by the company it keeps”
J. R. Firth, British Linguist, 1957

Male-Female

Verb tense

Country-Capital
Context-GPU: Combining local and global context

Simple Polya Urn model

Generalized Polya Urn model

**Context-GPU:** Combining local and global context

Context - Generalized Polya Urn (GPU) model

- **Local_context**: (i.e. windows-based embeddings)
- **Global_context**: (i.e. document-based embeddings)
**Context-GPU:** Combining local and global context

**Context - Generalized Polya Urn (GPU) model**

**Local_context** ($W_1$)
- windows-based embedding
  - (e.g. FastText)

**Global_context** ($W_2$)
- document-based embedding
  - (e.g. LSI – Latent Semantic Indexing)

**Enhancements**
- Improve word semantic
- Add external knowledge
- Simple and efficient process
Context-GPU

 [...] white_blood_cell [...] 
 [...] shortness_of_breath [...] 
 ...
Context-GPU

[...] white_blood_cell [...]  
[...] shortness_of_breath [...]  
...
Context-GPU

[...] white_blood_cell [...] 
[...] shortness_of_breath [...]
Context-GPU

[... white_blood_cell [...]]
[... shortness_of_breath [...]]

...
Context-GPU

[..] white_blood_cell [..]
[..] shortness_of_breath [..]

FastText

MedTagger

LSI

Coronary artery d.
Cardiac transp.
Cardiomyopathy
Pravachol 20 mg

Pregnancy Ultrasound
Postpartum hem...
Endometrial biop...

Congestive hear...
Pulmonary edema
Orthopnea
Nonischemic
Diastolic dysfunct...

Right coronary art.

ContextGPU
I2b2 dataset

- 1,243 de-identified discharge summaries
- 7,883 unique terms with “bag of words”
- 9,932 unique terms with “bag of phrases”
- Patient history, discharge, medications and treatments, etc.

“Community annotation experiment for ground truth generation for the i2b2 medication challenge”, Journal of the American Medical Informatics Association, Uzuner et al. 2010
<table>
<thead>
<tr>
<th>Qualitative assessment</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>LDA</th>
<th>Contex-GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemotherapy</td>
<td>congestive heart failure</td>
</tr>
<tr>
<td>dilantin</td>
<td>diuresis</td>
</tr>
<tr>
<td>oncology</td>
<td>ejection fraction</td>
</tr>
<tr>
<td>xrt</td>
<td>approximately</td>
</tr>
<tr>
<td>oncologist</td>
<td>felt</td>
</tr>
<tr>
<td>cycle</td>
<td>orthopnea</td>
</tr>
<tr>
<td>breast cancer</td>
<td>digoxin</td>
</tr>
<tr>
<td>cancer</td>
<td>dyspnea</td>
</tr>
<tr>
<td>seizure</td>
<td>weight</td>
</tr>
<tr>
<td>left breast</td>
<td>pressure</td>
</tr>
<tr>
<td>pregnancy</td>
<td>coronary artery disease</td>
</tr>
<tr>
<td>ultrasound</td>
<td>congestive heart failure</td>
</tr>
<tr>
<td>postpartum hemorrhage</td>
<td>pulmonary edema</td>
</tr>
<tr>
<td>endometrial biopsy</td>
<td>orthopnea</td>
</tr>
<tr>
<td>total abdominal hysterectomy</td>
<td>nonischemic</td>
</tr>
<tr>
<td>postpartum</td>
<td>diastolic dysfunction</td>
</tr>
<tr>
<td>vomiting</td>
<td>cardiomyopathy</td>
</tr>
<tr>
<td>salpingo oophorectomy</td>
<td>heart failure</td>
</tr>
<tr>
<td>physical examination</td>
<td>shortness of breath</td>
</tr>
<tr>
<td>fibroid</td>
<td>cardiac catheterization</td>
</tr>
<tr>
<td></td>
<td>atrial fibrillation</td>
</tr>
</tbody>
</table>
## Qualitative assessment

<table>
<thead>
<tr>
<th>LDA</th>
<th>Context-GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemotherapy</td>
<td>congestive heart failure</td>
</tr>
<tr>
<td>dilantin</td>
<td>diuresis</td>
</tr>
<tr>
<td>oncology</td>
<td>ejection fraction</td>
</tr>
<tr>
<td>xrt</td>
<td>approximately</td>
</tr>
<tr>
<td>oncologist</td>
<td>felt</td>
</tr>
<tr>
<td>cycle</td>
<td>orthopnea</td>
</tr>
<tr>
<td>breast cancer</td>
<td>digoxin</td>
</tr>
<tr>
<td>cancer</td>
<td>dyspnea</td>
</tr>
<tr>
<td>seizure</td>
<td>weight</td>
</tr>
<tr>
<td>left breast</td>
<td>pressure</td>
</tr>
<tr>
<td></td>
<td>pregnancy</td>
</tr>
<tr>
<td></td>
<td>ultrasound</td>
</tr>
<tr>
<td>postpartum hemorrhage</td>
<td>coronary artery disease</td>
</tr>
<tr>
<td>endometrial biopsy</td>
<td>cardiac transplant</td>
</tr>
<tr>
<td>total abdominal hysterectomy</td>
<td>cardiomyopathy</td>
</tr>
<tr>
<td>postpartum</td>
<td>right coronary artery</td>
</tr>
<tr>
<td>vomiting</td>
<td>pravachol 20 mg</td>
</tr>
<tr>
<td>salpingo oophorectomy</td>
<td>paroxysmal atrial fibrillation</td>
</tr>
<tr>
<td>physical examination</td>
<td>cyclosporine</td>
</tr>
<tr>
<td>fibroid</td>
<td>herpes zoster</td>
</tr>
<tr>
<td></td>
<td>fenofibrate tricor</td>
</tr>
<tr>
<td></td>
<td>right coronary artery</td>
</tr>
<tr>
<td></td>
<td>congestive heart failure</td>
</tr>
<tr>
<td></td>
<td>pulmonary edema</td>
</tr>
<tr>
<td></td>
<td>orthopnea</td>
</tr>
<tr>
<td></td>
<td>nonischemic</td>
</tr>
<tr>
<td></td>
<td>diastolic dysfunction</td>
</tr>
<tr>
<td></td>
<td>cardiomyopathy</td>
</tr>
<tr>
<td></td>
<td>heart failure</td>
</tr>
<tr>
<td></td>
<td>shortness of breath</td>
</tr>
<tr>
<td></td>
<td>cardiac catheterization</td>
</tr>
<tr>
<td></td>
<td>atrial fibrillation</td>
</tr>
<tr>
<td>Qualitative assessment</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemotherapy</td>
</tr>
<tr>
<td>dilantin</td>
</tr>
<tr>
<td>oncology</td>
</tr>
<tr>
<td>xrt</td>
</tr>
<tr>
<td>oncologist</td>
</tr>
<tr>
<td>cycle</td>
</tr>
<tr>
<td>breast cancer</td>
</tr>
<tr>
<td>cancer</td>
</tr>
<tr>
<td>seizure</td>
</tr>
<tr>
<td>left breast</td>
</tr>
<tr>
<td>fever</td>
</tr>
<tr>
<td>urinalysis</td>
</tr>
<tr>
<td>culture</td>
</tr>
<tr>
<td>bacteria</td>
</tr>
<tr>
<td>white blood cell</td>
</tr>
<tr>
<td>polys</td>
</tr>
<tr>
<td>urinary tract infection</td>
</tr>
<tr>
<td>band</td>
</tr>
<tr>
<td>fluid</td>
</tr>
<tr>
<td>white</td>
</tr>
<tr>
<td>congestive heart failure</td>
</tr>
<tr>
<td>diuresis</td>
</tr>
<tr>
<td>ejection fraction</td>
</tr>
<tr>
<td>approximately</td>
</tr>
<tr>
<td>fall</td>
</tr>
<tr>
<td>orthopnea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context-GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>pregnancy</td>
</tr>
<tr>
<td>ultrasound</td>
</tr>
<tr>
<td>postpartum hemorrhage</td>
</tr>
<tr>
<td>endometrial biopsy</td>
</tr>
<tr>
<td>total abdominal hysterectomy</td>
</tr>
<tr>
<td>postpartum</td>
</tr>
<tr>
<td>vomiting</td>
</tr>
<tr>
<td>salpingo oophorectomy</td>
</tr>
<tr>
<td>physical examination</td>
</tr>
<tr>
<td>fibroid</td>
</tr>
<tr>
<td>coronary artery disease</td>
</tr>
<tr>
<td>cardiac transplant</td>
</tr>
<tr>
<td>cardiomyopathy</td>
</tr>
<tr>
<td>right coronary artery</td>
</tr>
<tr>
<td>pravachol 20 mg</td>
</tr>
<tr>
<td>paroxysmal atrial fibrillation</td>
</tr>
<tr>
<td>cyclosporine</td>
</tr>
<tr>
<td>herpes zoster</td>
</tr>
<tr>
<td>fenofibrate tricor</td>
</tr>
<tr>
<td>right coronary artery</td>
</tr>
<tr>
<td>congestive heart failure</td>
</tr>
<tr>
<td>pulmonary edema</td>
</tr>
<tr>
<td>orthopnea</td>
</tr>
<tr>
<td>nonschismic</td>
</tr>
<tr>
<td>diastolic dysfunction</td>
</tr>
<tr>
<td>cardiomyopathy</td>
</tr>
<tr>
<td>heart failure</td>
</tr>
<tr>
<td>shortness of breath</td>
</tr>
<tr>
<td>cardiac catheterization</td>
</tr>
<tr>
<td>atrial fibrillation</td>
</tr>
</tbody>
</table>
Context-GPU - Performance
Conclusions

• **Scale up** to bigger dataset (from the i2b2 to the Mimic III dataset) or documents in a new domain (e.g. legal documents)

• Substitute more context-depend representations using Contextualized Language Models within the inference process. In addition, **WordPiece tokenizer** combining sub-tokens could better fit the topical phrase encoding.

• **Promotion of medical entities** by means of a biomedical NER (e.g. SciSpacy).

• Not only topic, but **medical coherence**
Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies

Gabriele Pergola†, Elena Kochkina§, Lin Gui †, Maria Liakata* and Yulan He†

†University of Warwick
§Alan Turing Institute
*Queen Mary University Of London

European Chapter of the Association for Computational Linguistics (EACL)

2021
Outline

• Brief Introduction to the QA task
• Why is Biomedical QA hard?
• Transfer Learning
• Biomedical Entity-Aware Masking (BEM) Strategy
• Experimental assessment
• Conclusion and Future Work
Why Biomedical Question Answering?

Find me an answer among scientific papers of the current literature

I need more information on my health problem

Find me relevant information on this issue!
Question-Answering

**Questions:** Who patented a steam engine in 1781?

**Answer:** James Watt
Questions: Who patented a steam engine in 1781?

Passage: In 1781 James Watt patented a steam engine that produced continuous rotary motion. Watt’s ten-horsepower engines enabled a wide range of manufacturing machinery to be powered. The engines could be sited anywhere that water and coal or wood fuel could be obtained. By 1883, engines that could provide 10,000 hp had become feasible. The stationary steam engine was a key component of the Industrial Revolution, allowing factories to locate where water power was unavailable. The atmospheric engines of Newcomen and Watt were large compared to the amount of power they produced, but high pressure steam engines were light enough to be applied to vehicles such as traction engines and the railway locomotives.
The mean incubation period was 5.6 days (95% CI: 4.4, 7.4) when excluding Wuhan residents—slightly larger than the estimate without right truncation. The mean estimate for illness onset to hospital admission was 9.7 days (95% CI: 5.4, 17.0) for living cases and 6.6 days (95% CI: 5.2, 8.8) for deceased cases, with the former nearly 2.5 times the length of its untruncated version. Illness onset to death and hospital admission to death were likewise longer than their non-truncated counterparts, at 20.2 days (95% CI: 15.1, 29.5) and 13.0 days (95% CI: 8.7, 20.9), respectively.

Questions: What is the incubation period of the virus?
Why is Biomedical QA hard?

- **COMPOSITIONAL SEMANTIC**

  ‘spike protein sequence’ and ‘white blood cell’.
Why is Biomedical QA hard?

- **COMPOSITIONAL SEMANTIC**  
  ‘spike protein sequence’ and ‘white blood cell’.

- **TECHNICAL JARGON**  
  ‘basophilia’, ‘synechococcus elongatus’.
Why is Biomedical QA hard?

• **COMPOSITIONAL SEMANTIC**
  
  ‘spike protein sequence’ and ‘white blood cell’.

• **TECHNICAL JARGON**
  
  ‘basophilia’, ‘synechococcus elongatus’.

• **ABBREVIATIONS**
  
  S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’
  - ‘systolic blood pressure’
Why is Biomedical QA hard?

- **COMPOSITIONAL SEMANTIC**
  
  ‘spike protein sequence’ and ‘white blood cell’.

- **TECHNICAL JARGON**
  
  ‘basophilia’, ‘synechococcus elongatus’.

- **ABBREVIATIONS**
  
  S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’  
  - ‘systolic blood pressure’

- **POLYSEMY**
  
  ‘Inflammation’ up to 5 different meanings1.

---

[1]: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Why is Biomedical QA hard?

- **COMPOSITIONAL SEMANTIC**
  - ‘spike protein sequence’ and ‘white blood cell’.

- **TECHNICAL JARGON**
  - ‘basophilia’, ‘synechococcus elongatus’.

- **ABBREVIATIONS**
  - S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’
    - ‘systolic blood pressure’

- **POLYSEMY**
  - ‘Inflammation’ up to 5 different meanings\(^1\).

- **LACK OF STRUCTURE**

---

\[^1\]: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Why is Biomedical QA hard?

- **COMPOSITIONAL SEMANTIC**
  - ‘spike protein sequence’ and ‘white blood cell’.

- **TECHNICAL JARGON**
  - ‘basophilia’, ‘synechococcus elongatus’.

- **ABBREVIATIONS**
  - S.B.P. expanded both as: - ‘spontaneous bacterial peritonitis’
    - ‘systolic blood pressure’

- **POLYSEMY**
  - ‘Inflammation’ up to 5 different meanings\(^1\).

- **LACK OF STRUCTURE**

- **DATA AVAILABILITY**
  - CovidQA \([1]\): 127 pairs
  - BioASQ \([2]\): ~2000 pairs
  - SQuAD \([3]\): ~150.000 pairs

---

\[1\]: “Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies”, Pisanelli et al. 2004
Transfer Learning: Sequential Adaptation

Pre-Training

Wikipedia
BookCorpus
...

→

Masked Language Model (MLM)
Transfer Learning: **Sequential Adaptation**

Pre-Training

- Wikipedia
- BookCorpus
- ...

**Masked Language Model (MLM)**

Sentence:

```
The dog was chewing on a [MASK].
```

Mask 1 Predictions:

- 45.2% bone
- 30.1% stick
- 15.3% toy
- 9.4% shoe
Transfer Learning: **Sequential Adaptation**

**Pre-Training**

- Wikipedia
- BookCorpus
- ...

→

**Fine-Tuning**

- MLM

→

**Questions:** What is the incubation period of the virus?

**Passage:** The mean incubation period was 5.6 days (95% CI: 4.4, 7.4) when excluding Wuhan residents—slightly larger than the estimate without right truncation. The mean estimate for illness onset to hospital admission was 9.7 days (95% CI: 5.4, 17.0) for living cases and 6.6 days (95% CI: 5.2, 8.8) for deceased cases, with the former nearly 2.5 times the length of its untruncated version. Illness onset to death and hospital admission to death were likewise longer than their non-truncated counterparts, at 20.2 days (95% CI: 15.1, 25.3) and 13.0 days (95% CI: 8.7, 20.9), respectively.

**BioASQ**

**CovidQA**

...
Transfer Learning: **Sequential Adaptation**

**Pre-Training**
- Wikipedia
- BookCorpus
- ...

**Fine-Tuning**
- **Domain Adaptation**
- **Task Adaptation**
- Question Answering

**MLM**

**BioBERT**
**RoBERTa-BioMED**

[1]: “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”, Lee et al., Bioinformatics 2019
[2]: “Don’t stop pretraining: Adapt language models to domains and tasks”, Gururangan et al., ACL20
Transfer Learning: **Sequential Adaptation**

**Pre-Training**

- Wikipedia
- BookCorpus
- ...

→ **MLM** →

**Fine-Tuning**

→ **Question Answering**

*Example*: Use of pre-trained models for fine-tuning.
Transfer Learning: **Sequential Adaptation**

Pre-Training → MLM → Fine-Tuning

- Wikipedia
- BookCorpus

Fine-Tuning

Question Answering
Transfer Learning: **Sequential Adaptation**

Pre-Training → **MLM** → Fine-Tuning

**Attention On Bio-Medical Entities**

*Input*: Wikipedia, BookCorpus, …

*Output*: Question Answering
Attention On Medical Entities
Masked Language Models

Masked Language Model
Generates word representations that can be used to predict the missing tokens of an input text.

```
the man went to the [MASK] to buy a [MASK] of wine
```

[1]: “BERT: Pre-training of deep bidirectional transformers for language understanding”, Devlin et al., NAACL19
Masked Language Models

Masked Language Model
Generates word representations that can be used to predict the missing tokens of an input text.

![Masked Language Model Example]

BERT
Pre-trained language model (LM) using multi-layer bidirectional transformer networks to encode contextualised language representations.

BERT’s Masking Strategy
15% tokens randomly chosen:
- 80% replaced with [MASK]
- 10% randomly swapped
- 10% kept the same
Biomedical Entity-Aware Masking Strategy

BEM masks a proportion of the medical entities in text. Thus, realigning the word representation to predict the missing medical entities.

Patients with diabetes (HR 1.59) were more likely to reach to the composite endpoints than those without.
Patients with diabetes (HR 1.59) were more likely to reach to the composite endpoints than those without.
Biomedical Entity-Aware Masking Strategy

BEM masks a proportion of the medical entities in text. Thus, realigning the word representation to predict the missing medical entities.

Patients diabetes HR

[MASK] with [MASK] ([MASK] 1.59) were more likely to reach to the [MASK] [MASK] than those without.

composite endpoints
Patients with diabetes (HR 1.59) were more likely to reach the composite endpoints than those without.
Transfer Learning: **Sequential Adaptation**

### Pre-Training
- Large Corpora
- Pre-Training Masked Language Model

### Fine-Tuning
- Biomedical Entities
- BEM Fine-Tuning
- QA Fine-Tuning

#### MLM
- Patients
- Diabetes
- HR
- Composite Endpoints
- Covid-19
- Coronary Pneumonia

---

**Wikipedia**
**BookCorpus**

---

**Transfer Learning:** Sequential Adaptation
Results: CovidQA and BioASQ dataset


• CovidQA [3] is a manually curated dataset based on the AI2’s COVID-19 Open Re-search dataset, and consists of 127 question-answer pairs with 27 questions and 85 unique related articles.

What is the **incubation period** of the virus?  
What is the length of viral **shedding** after illness onset?  
What is the **incubation period** across different age groups?  
What is the proportion of patients who were **asymptomatic**?  
What is the **asymptomatic transmission** during incubation?

[1]: “An overview of the BioASQ largescale biomedical semantic indexing and question answering competition”, Tsatsaronis et al., BMC Bioinformatics 2015  
[3]: “Rapidly bootstrapping a question answering dataset for COVID-19” R. Tang et al. 2020
Results: CovidQA and BioASQ dataset


- **CovidQA** [3] is a manually curated dataset based on the AI2’s COVID-19 Open Re-search dataset, and consists of 127 question-answer pairs with 27 questions and 85 unique related articles.

What is the RR for severe infection in COVID-19 patients with hypertension?
What is the HR for severe infection in COVID-19 patients with hypertension?
What is the OR for severe infection in COVID-19 patients with hypertension?

[1]: “An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition”, Tsatsaronis et al., BMC Bioinformatics 2015
[3]: “Rapidly bootstrapping a question answering dataset for COVID-19” R. Tang et al. 2020
Results: CovidQA and BioASQ dataset


- **CovidQA** [3] is a manually curated dataset based on the AI2’s COVID-19 Open Re-search dataset, and consists of 127 question-answer pairs with 27 questions and 85 unique related articles.

- **BioASQ** is one of the larger biomedical QA datasets available with over 2000 question-answer pairs. To use it within the extractive questions-answering framework, we convert the questions into the SQuAD dataset format.

[1]: “An overview of the BioASQ largescale biomedical semantic indexing and question answering competition”, Tsatsaronis et al., BMC Bioinformatics 2015
[3]: “Rapidly bootstrapping a question answering dataset for COVID-19” R. Tang et al. 2020
Results: CovidQA and BioASQ dataset


• BioASQ is one of the larger biomedical QA datasets available with over 2000 question-answer pairs. To use it within the extractive questions-answering framework, we convert the questions into the SQuAD dataset format.

What does the pembrolizumab companion diagnostic test assess?
What is the combined effect of Nfat and miR-25?
Which genomic positions are preferentially selected for transposon insertion?
Is Baloxavir effective for influenza?
## Results: QA Metrics

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th>CovidQA P@1</th>
<th>CovidQA R@3</th>
<th>CovidQA MRR</th>
<th>BioASQ 7b SAcc</th>
<th>BioASQ 7b LAcc</th>
<th>BioASQ 7b MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT</td>
<td>0.081*</td>
<td>0.117*</td>
<td>0.159*</td>
<td>0.012</td>
<td>0.032</td>
<td>0.027</td>
</tr>
<tr>
<td>2</td>
<td>BERT + BioASQ</td>
<td>0.125</td>
<td>0.177</td>
<td>0.206</td>
<td>0.226</td>
<td>0.317</td>
<td>0.262</td>
</tr>
<tr>
<td>3</td>
<td>BERT + STM + BioASQ</td>
<td>0.132</td>
<td>0.195</td>
<td>0.218</td>
<td>0.233</td>
<td>0.325</td>
<td>0.265</td>
</tr>
<tr>
<td>4</td>
<td>BERT + BEM + BioASQ</td>
<td>0.145</td>
<td>0.278</td>
<td>0.269</td>
<td>0.241</td>
<td>0.341</td>
<td>0.288</td>
</tr>
<tr>
<td>5</td>
<td>RoBERTa</td>
<td>0.068</td>
<td>0.115</td>
<td>0.122</td>
<td>0.023</td>
<td>0.041</td>
<td>0.036</td>
</tr>
<tr>
<td>6</td>
<td>RoBERTa + BioASQ</td>
<td>0.106</td>
<td>0.155</td>
<td>0.178</td>
<td>0.278</td>
<td>0.324</td>
<td>0.294</td>
</tr>
<tr>
<td>7</td>
<td>RoBERTa + STM + BioASQ</td>
<td>0.112</td>
<td>0.167</td>
<td>0.194</td>
<td>0.282</td>
<td>0.333</td>
<td>0.300</td>
</tr>
<tr>
<td>8</td>
<td>RoBERTa + BEM + BioASQ</td>
<td>0.125</td>
<td>0.198</td>
<td>0.236</td>
<td>0.323</td>
<td>0.374</td>
<td>0.325</td>
</tr>
<tr>
<td>9</td>
<td>RoBERTa-Biomed</td>
<td>0.104</td>
<td>0.163</td>
<td>0.192</td>
<td>0.028</td>
<td>0.044</td>
<td>0.037</td>
</tr>
<tr>
<td>10</td>
<td>RoBERTa-Biomed + BioASQ</td>
<td>0.128</td>
<td>0.355</td>
<td>0.315</td>
<td>0.415</td>
<td>0.398</td>
<td>0.376</td>
</tr>
<tr>
<td>11</td>
<td>RoBERTa-Biomed + STM + BioASQ</td>
<td>0.136</td>
<td>0.364</td>
<td>0.321</td>
<td>0.423</td>
<td>0.410</td>
<td>0.397</td>
</tr>
<tr>
<td>12</td>
<td>RoBERTa-Biomed + BEM + BioASQ</td>
<td>0.143</td>
<td>0.386</td>
<td>0.347</td>
<td>0.435</td>
<td>0.443</td>
<td>0.398</td>
</tr>
<tr>
<td>13</td>
<td>BioBERT</td>
<td>0.097*</td>
<td>0.142*</td>
<td>0.170*</td>
<td>0.031</td>
<td>0.046</td>
<td>0.039</td>
</tr>
<tr>
<td>14</td>
<td>BioBERT + BioASQ</td>
<td>0.166</td>
<td>0.419</td>
<td>0.348</td>
<td>0.410†</td>
<td>0.474†</td>
<td>0.409†</td>
</tr>
<tr>
<td>15</td>
<td>BioBERT + STM + BioASQ</td>
<td>0.172</td>
<td>0.432</td>
<td>0.385</td>
<td>0.418</td>
<td>0.482</td>
<td>0.416</td>
</tr>
<tr>
<td>16</td>
<td>BioBERT + BEM + BioASQ</td>
<td>0.179</td>
<td><strong>0.458</strong></td>
<td><strong>0.391</strong></td>
<td>0.421</td>
<td><strong>0.497</strong></td>
<td><strong>0.434</strong></td>
</tr>
<tr>
<td>17</td>
<td>T5 LM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>T5 LM + MS–MARCO</td>
<td><strong>0.282</strong></td>
<td><strong>0.404</strong></td>
<td><strong>0.415</strong></td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
## Results: QA Metrics

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th>CovidQA P@1</th>
<th>CovidQA R@3</th>
<th>CovidQA MRR</th>
<th>BioASQ 7b SAcc</th>
<th>BioASQ 7b LAcc</th>
<th>BioASQ 7b MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT</td>
<td>0.081*</td>
<td>0.117*</td>
<td>0.159*</td>
<td>0.012</td>
<td>0.032</td>
<td>0.027</td>
</tr>
<tr>
<td>2</td>
<td>+ BioASQ</td>
<td>0.125</td>
<td>0.177</td>
<td>0.206</td>
<td>0.226</td>
<td>0.317</td>
<td>0.262</td>
</tr>
<tr>
<td>3</td>
<td>+ STM + BioASQ</td>
<td>0.132</td>
<td>0.195</td>
<td>0.218</td>
<td>0.233</td>
<td>0.325</td>
<td>0.265</td>
</tr>
<tr>
<td>4</td>
<td>+ BEM + BioASQ</td>
<td>0.145</td>
<td>0.278</td>
<td>0.269</td>
<td>0.241</td>
<td>0.341</td>
<td>0.288</td>
</tr>
<tr>
<td>5</td>
<td>RoBERTa</td>
<td>0.068</td>
<td>0.115</td>
<td>0.122</td>
<td>0.023</td>
<td>0.041</td>
<td>0.036</td>
</tr>
<tr>
<td>6</td>
<td>+ BioASQ</td>
<td>0.106</td>
<td>0.155</td>
<td>0.178</td>
<td>0.278</td>
<td>0.324</td>
<td>0.294</td>
</tr>
<tr>
<td>7</td>
<td>+ STM + BioASQ</td>
<td>0.112</td>
<td>0.167</td>
<td>0.194</td>
<td>0.282</td>
<td>0.333</td>
<td>0.300</td>
</tr>
<tr>
<td>8</td>
<td>+ BEM + BioASQ</td>
<td>0.125</td>
<td>0.198</td>
<td>0.236</td>
<td>0.323</td>
<td>0.374</td>
<td>0.325</td>
</tr>
<tr>
<td>9</td>
<td>RoBERTa-Biomed</td>
<td>0.104</td>
<td>0.163</td>
<td>0.192</td>
<td>0.028</td>
<td>0.044</td>
<td>0.037</td>
</tr>
<tr>
<td>10</td>
<td>+ BioASQ</td>
<td>0.128</td>
<td>0.355</td>
<td>0.315</td>
<td>0.415</td>
<td>0.398</td>
<td>0.376</td>
</tr>
<tr>
<td>11</td>
<td>+ STM + BioASQ</td>
<td>0.136</td>
<td>0.364</td>
<td>0.321</td>
<td>0.423</td>
<td>0.410</td>
<td>0.397</td>
</tr>
<tr>
<td>12</td>
<td>+ BEM + BioASQ</td>
<td>0.143</td>
<td>0.386</td>
<td>0.347</td>
<td><strong>0.435</strong></td>
<td>0.443</td>
<td>0.398</td>
</tr>
<tr>
<td>13</td>
<td>BioBERT</td>
<td>0.097*</td>
<td>0.142*</td>
<td>0.170*</td>
<td>0.031</td>
<td>0.046</td>
<td>0.039</td>
</tr>
<tr>
<td>14</td>
<td>+ BioASQ</td>
<td>0.166</td>
<td>0.419</td>
<td>0.348</td>
<td><strong>0.410</strong>(†)</td>
<td><strong>0.474</strong>(†)</td>
<td><strong>0.409</strong>(†)</td>
</tr>
<tr>
<td>15</td>
<td>+ STM + BioASQ</td>
<td>0.172</td>
<td>0.432</td>
<td>0.385</td>
<td>0.418</td>
<td>0.482</td>
<td>0.416</td>
</tr>
<tr>
<td>16</td>
<td>+ BEM + BioASQ</td>
<td>0.179</td>
<td><strong>0.458</strong></td>
<td>0.391</td>
<td>0.421</td>
<td><strong>0.497</strong></td>
<td><strong>0.434</strong></td>
</tr>
<tr>
<td>17</td>
<td>T5 LM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>+ MS–MARCO</td>
<td><strong>0.282</strong>*</td>
<td><strong>0.404</strong>*</td>
<td><strong>0.415</strong>*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Results: QA Metrics

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th>CovidQA</th>
<th></th>
<th>BioASQ 7b</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P@1</td>
<td>R@3</td>
<td>MRR</td>
<td>SAcc</td>
</tr>
<tr>
<td>1</td>
<td>BERT</td>
<td>0.081*</td>
<td>0.117*</td>
<td>0.159*</td>
<td>0.012</td>
</tr>
<tr>
<td>2</td>
<td>+ BioASQ</td>
<td>0.125</td>
<td>0.177</td>
<td>0.206</td>
<td>0.226</td>
</tr>
<tr>
<td>3</td>
<td>+ STM + BioASQ</td>
<td>0.132</td>
<td>0.195</td>
<td>0.218</td>
<td>0.233</td>
</tr>
<tr>
<td>4</td>
<td>+ BEM + BioASQ</td>
<td>0.145</td>
<td>0.278</td>
<td>0.269</td>
<td>0.241</td>
</tr>
<tr>
<td>5</td>
<td>RoBERTa</td>
<td>0.068</td>
<td>0.115</td>
<td>0.122</td>
<td>0.023</td>
</tr>
<tr>
<td>6</td>
<td>+ BioASQ</td>
<td>0.106</td>
<td>0.155</td>
<td>0.178</td>
<td>0.278</td>
</tr>
<tr>
<td>7</td>
<td>+ STM + BioASQ</td>
<td>0.112</td>
<td>0.167</td>
<td>0.194</td>
<td>0.282</td>
</tr>
<tr>
<td>8</td>
<td>+ BEM + BioASQ</td>
<td>0.125</td>
<td>0.198</td>
<td>0.236</td>
<td>0.323</td>
</tr>
<tr>
<td>9</td>
<td>RoBERTa-Biomed</td>
<td>0.104</td>
<td>0.163</td>
<td>0.192</td>
<td>0.028</td>
</tr>
<tr>
<td>10</td>
<td>+ BioASQ</td>
<td>0.128</td>
<td>0.355</td>
<td>0.315</td>
<td>0.415</td>
</tr>
<tr>
<td>11</td>
<td>+ STM + BioASQ</td>
<td>0.136</td>
<td>0.364</td>
<td>0.321</td>
<td>0.423</td>
</tr>
<tr>
<td>12</td>
<td>+ BEM + BioASQ</td>
<td>0.143</td>
<td>0.386</td>
<td>0.347</td>
<td>0.435</td>
</tr>
<tr>
<td>13</td>
<td>BioBERT</td>
<td>0.097*</td>
<td>0.142*</td>
<td>0.170*</td>
<td>0.031</td>
</tr>
<tr>
<td>14</td>
<td>+ BioASQ</td>
<td>0.166</td>
<td>0.419</td>
<td>0.348</td>
<td>0.410†</td>
</tr>
<tr>
<td>15</td>
<td>+ STM + BioASQ</td>
<td>0.172</td>
<td>0.432</td>
<td>0.385</td>
<td>0.418</td>
</tr>
<tr>
<td>16</td>
<td>+ BEM + BioASQ</td>
<td>0.179</td>
<td>0.458</td>
<td>0.391</td>
<td>0.421</td>
</tr>
<tr>
<td>17</td>
<td>T5 LM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>+ MS–MARCO</td>
<td>0.282*</td>
<td>0.404*</td>
<td>0.415*</td>
<td>—</td>
</tr>
</tbody>
</table>
Results: Qualitative Results

<table>
<thead>
<tr>
<th>BERT with STM</th>
<th>BERT with BEM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is the OR for severe infection in COVID-19 patients with hypertension?</strong></td>
<td><strong>What is the OR for severe infection in COVID-19 patients with hypertension?</strong></td>
</tr>
<tr>
<td>- There were significant correlations between COVID-19 severity and [...], diabetes [OR=2.67], coronary heart disease [OR=2.85].</td>
<td>- There were significant correlations between COVID-19 severity and [...], diabetes [OR=2.67], coronary heart disease [OR=2.85].</td>
</tr>
<tr>
<td>- Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).</td>
<td>- Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).</td>
</tr>
</tbody>
</table>

| **What is the HR for severe infection in COVID-19 patients with hypertension?** | **What is the HR for severe infection in COVID-19 patients with hypertension?** |
| - - - - | - After adjusting for age and smoking status, patients with COPD (HR 2.681), diabetes (HR 1.59), and malignancy (HR 3.50) were more likely to reach to the composite endpoints than those without. |

| **What is the RR for severe infection in COVID-19 patients with hypertension?** | **What is the RR for severe infection in COVID-19 patients with hypertension?** |
| - - - - | - In univariate analyses, factors significantly associated with severe COVID-19 were male sex (14 studies; pooled RR=1.70, ..), hypertension (10 studies 2.74 ...),diabetes (11 studies ...), and CVD (..). |
## Results: Qualitative Results

### BERT with STM

**What is the OR for severe infection in COVID-19 patients with hypertension?**

- There were significant correlations between COVID-19 severity and [...], diabetes [OR=2.67], coronary heart disease [OR=2.85].
- Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).

**What is the HR for severe infection in COVID-19 patients with hypertension?**

- - - -

**What is the RR for severe infection in COVID-19 patients with hypertension?**

- - - -

### BERT with BEM

**What is the OR for severe infection in COVID-19 patients with hypertension?**

- There were significant correlations between COVID-19 severity and [...], diabetes [OR=2.67], coronary heart disease [OR=2.85].
- Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).

**What is the HR for severe infection in COVID-19 patients with hypertension?**

- After adjusting for age and smoking status, patients with COPD (HR 2.681), diabetes (HR 1.59), and malignancy (HR 3.50) were more likely to reach to the composite endpoints than those without.

**What is the RR for severe infection in COVID-19 patients with hypertension?**

- In univariate analyses, factors significantly associated with severe COVID-19 were male sex (14 studies; pooled RR=1.70, ..), hypertension (10 studies 2.74 ..., diabetes (11 studies ...), and CVD (..).
Results: Qualitative Results

<table>
<thead>
<tr>
<th>BERT with STM</th>
<th>BERT with BEM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is the OR for severe infection in COVID-19 patients with hypertension?</strong></td>
<td></td>
</tr>
</tbody>
</table>

- There were significant correlations between COVID-19 severity and [..], diabetes [OR=2.67], coronary heart disease [OR=2.85].
- Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).

| **What is the HR for severe infection in COVID-19 patients with hypertension?** |

- - - -

| **What is the RR for severe infection in COVID-19 patients with hypertension?** |

- - - -

- After adjusting for age and smoking status, patients with COPD (HR 2.681), diabetes (HR 1.59), and malignancy (HR 3.50) were more likely to reach to the composite endpoints than those without.

- In univariate analyses, factors significantly associated with severe COVID-19 were male sex (14 studies; pooled RR=1.70, ..), hypertension (10 studies 2.74 ...), diabetes (11 studies ...), and CVD (..).
Results: Qualitative Results

<table>
<thead>
<tr>
<th>BERT with STM</th>
<th>BERT with BEM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is the OR for severe infection in COVID-19 patients with hypertension?</strong></td>
<td></td>
</tr>
<tr>
<td>- There were significant correlations between COVID-19 severity and [...] diab[etes [OR=2.67], coronary heart disease [OR=2.85].</td>
<td></td>
</tr>
<tr>
<td>- Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).</td>
<td></td>
</tr>
</tbody>
</table>

| **What is the HR for severe infection in COVID-19 patients with hypertension?** |
| - After adjusting for age and smoking status, patients with COPD (HR 2.681), diabetes (HR 1.59), and malignancy (HR 3.50) were more likely to reach to the composite endpoints than those without. |

| **What is the RR for severe infection in COVID-19 patients with hypertension?** |
| - In univariate analyses, factors significantly associated with severe COVID-19 were male sex (14 studies; pooled RR=1.70, ..), hypertension (10 studies 2.74 ...), diabetes (11 studies ...), and CVD (..). |
Results: Perplexity
Conclusions

Open questions:

• It remains worth investigating what would be the best strategy in choosing the medical entities to mask depending on their *types* (e.g. UMLS categories), and in what proportion doing so.

• Language model pre-training can capture world knowledge by storing it implicitly in the network parameters. More structured knowledge can be integrated to drive the masking process, adopting, for instance, a *knowledge graph* of the medical entities’ dependences.

• First step towards the integration of topics for a *Topic-Aware Masking Strategy*.

• Assess whether a similar technique would be applicable in other *low-resources domains*, or to drive the model towards *less represented instances* (mitigating the inherent bias).