Concept Wikification for COVID-19

Panagiotis Lymperopoulos
Tufts University
Medford, MA 02155
plympe01@tufts.edu

Haoling Qiu, Bonan Min
Raytheon BBN Technologies
Cambridge, MA 02138
{haoling.qiu,bonan.min}@
@raytheon.com

Abstract

Understanding scientific articles related to COVID-19 requires broad knowledge about concepts such as symptoms, diseases and medicine. Given the very large and ever-growing scientific articles related to COVID-19, it is a daunting task even for experts to recognize the large set of concepts mentioned in these articles. In this paper, we address the problem of concept wikification for COVID-19, which is to automatically recognize mentions of concepts related to COVID-19 in text and resolve them into Wikipedia titles. We develop an approach to curate a COVID-19 concept wikification dataset by mining Wikipedia text and the associated intra-Wikipedia links. We also develop an end-to-end system for concept wikification for COVID-19. Preliminary experiments show very encouraging results. Our dataset, code and pre-trained model are available at github.com/panlybero/Covid19_wikification.

1 Introduction

Recently, research articles about COVID-19 are published at overwhelming scale and speed. Understanding these articles is crucial for combating the COVID-19 pandemic. For example, the following sentence describes the transmission of coronavirus:

S1: *The interaction of the* **coronavirus spike protein** *with its complementary* **cell receptor** *is central in determining the* **tissue tropism**, **infectivity**, *and* **species range** *of the released virus. Coronaviruses mainly target* **epithelial cells**. *They are transmitted from one host to another host, depending on the coronavirus species, by either an* **aerosol**, **fomite**, *or* **fecal-oral** *route.*

There are a few concepts (in bold) that are central to understanding this sentence. Understanding these concepts requires biomedical expertise. The sheer volume of new information makes it a daunting task even for infectious disease experts to recognize the ever-growing set of concepts mentioned in the literature. This necessitates developing a concept recognition and linking system that can automatically tag mentions of these concepts in text and resolve them into Wikipedia titles.

We call this task *concept wikification for COVID-19*. It is an extension of the classic wikification (Roth et al., 2014) problem with a few new challenges introduced: first, what concepts are related to COVID-19? We need an efficient approach to identify these concepts at scale. Second, there is no labeled dataset for training such a wikifier. Therefore, we need an effective way to construct a labeled dataset timely, to assist researchers to develop concept wikifiers in time to combat the COVID-19 pandemic. Third, the concept wikifier has to perform both concept mention detection and linking in an end-to-end fashion, since there is no mention extraction algorithm for COVID-19 concepts yet. This is unlike classic wikification for which many Named Entity Recognition (NER) or mention detection systems are available.

In this paper, we address all these challenges and develop a practical solution for COVID-19 concept wikification. We first develop an approach to automatically discover concepts related to COVID-19 by exploring Wikipedia, starting from the Wikipedia page for COVID-19. We then leverage Wikipedia text and intra-Wikipedia links to automatically harvest a large dataset in which mentions are labeled with their Wikipedia titles. Using this dataset, we develop an end-to-end system for concept wikification for COVID-19. Preliminary experiments show encouraging results.

Our contributions are summarized as follows:

- We developed an approach to automatically identify 7,238 concepts related to COVID-19.
- We curated a large labeled dataset for training

concept wikifiers for these concepts. We made the dataset available to the public.

- We developed an end-to-end system for COVID-19 concept wikification. Preliminary results are encouraging.

2 Related Work

Entity Linking (EL) (Hachey et al., 2013; Guo et al., 2013; Yamada et al., 2016, 2017) or Wikification (Roth et al., 2014; Ratinov et al., 2011) is a reference resolution task, for which the goal is to identify entity mentions in text and resolve each mention into an entry in the reference knowledge base (KB) (for Wikification, the entry is a Wikipedia title). Most existing EL systems focus on entities such as persons, organizations and geo-political entities, and use NER or mention extraction systems (Ratinov and Roth, 2009) to identify candidate mentions. To perform disambiguation to Wikipedia titles, systems (Ratinov et al., 2011; Lin et al., 2017; Nguyen et al., 2016) often use lexical, syntactic and semantic features.

Recently, there has been growing interest in modeling the EL stages (e.g., mention detection, candidate generation, entity disambiguation) in a joint modeling framework such as a graphic model (Durrett and Klein, 2014) or a neural architecture (Nguyen et al., 2016; Kolitsas et al., 2018). (Broscheit, 2019) showed that a simple BERT-based token classification model can perform surprisingly well at end-to-end EL.

Entity linking research has benefit greatly from large-scale annotated datasets such as the CoNLL03/AIDA dataset (Hoffart et al., 2011). There are also approaches for specialized domains such as the bio-medical domain (Zheng et al., 2015), or in a cross-lingual setting (Sil et al., 2018; Pan et al., 2017).

This work differs from previous work in that we aim at recognizing and linking concepts related to COVID-19, to help timely and more comprehensive understanding of the concepts in the large amount of COVID-19 related scientific articles.

3 Building a Dataset for COVID-19 Concept Wikification

The goals are two-fold: first, we aim at finding concepts that are related to COVID-19. We also avoid concepts such as persons, organizations or geo-political entities (GPE), since they are covered in existing datasets. Second, we aim at harvesting text in which mentions of these concepts are labeled with their reference Wikipedia title. These labeled texts can be used to train a concept wikifier.

We leverage Wikipedia to construct this dataset because it has broad coverage, is frequently updated by its editors and has plenty of text and intra-page links that support these goals.

3.1 Finding Concepts Relevant to COVID-19

We repeat the following steps until we find a sufficient number of concepts relevant to COVID-19.

- Step 1: Start with known relevant pages $\mathcal{S}$ about COVID-19: parse each page $p \in \mathcal{S}$, and then find all concepts $\mathcal{C}$ mentioned in p.

- Step 2: For each $c \in \mathcal{C}$, fetch the title page p' for c and add p' into $\mathcal{S}$.

In a nutshell, we iteratively grow the set of COVID-19 relevant concepts by a breadth-first-search (BFS), starting at the Wikipedia entry for COVID-19 [1] and expanding outward by following links included in that page. We run BFS with a maximum depth of 2, to avoid exploding the search space. This results in a total of 11,795 concepts. The process is illustrated in Figure 1.

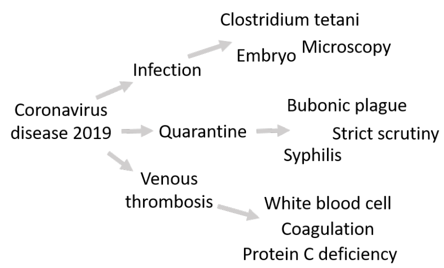

Figure 1: Sample search tree generated by a BFS starting from the Wikipedia entry for Covid-19. While only a branching factor of 3 is shown for illustration, the data collection procedure did not enforce any such limit.

Data collection from wikipedia is a noisy process, as even two iterations of BFS can reach pages that are only tangentially related to COVID-19. As a result, a few filters and post processing steps were applied to the raw data to narrow down the collection to relevant entities. First, we remove very rare concepts from the dataset, as these mostly consist of outliers. Second, we remove concepts of types such as cities and countries, that are covered extensively in existing datasets. [2]. We also used a list of

[1] Our initial $\mathcal{S}$ contains the single Wikipedia page about COVID-19: https://en.wikipedia.org/wiki/Coronavirus_disease_2019.

[2] The concept types are extracted from WikiData: https://www.wikidata.org/wiki.

cities and countries harvested from the GeoLite2 database [3], to prune entries that have no types in WikiData. Third, a small amount of manual effort is applied to select concept types that are related to COVID-19. We then keep the concepts whose types pass through these filters.

3.2 Harvesting Labeled Mentions for Concepts

To enable training a concept wikifier, we need to harvest labeled instances where each mention (text span) of a concept is labeled with a ground-truth Wikipedia title.

We generate labeled mentions for a concept by finding its occurrences in Wikipedia pages, via embedded intra-Wikipedia links that point to the concept. These intra-Wikipedia links are provided by Wikipedia editors. Since there are an overwhelming amount of instances for concepts such as *protein* or *coronavirus*, we cap the maximum numbers of sentences per concept to 50.

Identifying ambiguous mention strings A mention string m could be trivially resolved to its reference concept c if m only has one match in the vocabulary of concepts, since one could simply do string matching to resolve m to c. It is not uncommon for a scientific term (e.g., *SARS-CoV-2*) to always link to a unique wikipedia page, as it tends to be defined without ambiguity.

Based on the raw dataset consists of all mentions of all concepts, we calculate the probability of m being annotated as resolving to c as $p(c|m)$, which is the count of m labeled as c divided by the total count of m. To make a more challenging dataset for training our neural wikifier, we remove all instances of m where $p(c|m) < 0.9$ since these mention strings can be trivially resolved.

Table 1 shows the statistics of the resulting dataset. Example concept types, concepts and sentences are shown in Table 2.

Dataset	# of sentences	# of concepts
Full	108,868	7,238
Ambiguous	61,020	2,636

Table 1: Statistics of the COVID-19 concept linking dataset. The *ambiguous* dataset only includes ambiguous mentions for which $p(c|m) < 0.9$, while the *full* dataset includes all mentions for all concepts.

[3] This list is available at `github.com/panlybero/Covid19_wikification`

4 End-to-End Concept Wikification

We develop a hybrid, pipelined system consists of the following two steps. Given an input sentence, it first identifies concept strings (e.g., *SARS-CoV-2*) that are unambiguous and assigns their types using a majority-class classifier. It then applies a SciBERT (Beltagy et al., 2019)-based neural wikifier to assign types to other, more ambiguous concept mention strings.

Marjority-class classifier We first curate a list of unambiguous concept mention strings by finding mention string m such that $p(c|m) < 0.9$ for some c. It is trivial to resolve these mention strings into the corresponding concept. We perform string matching using the list of unambiguous mention strings, and simply resolve each match to its majority-class concept, which is the most frequent concept it was tagged in the dataset.

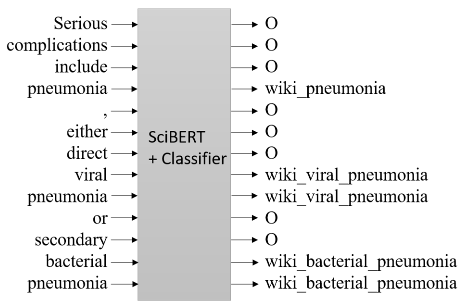

Figure 2: The neural SciBERT-based wikifier. It takes a sequence of words as input and outputs a sequence of labels corresponding to wikipedia titles or *Other* (O).

Neural SciBERT-based wikifier Inspired by (Broscheit, 2019) which showed that a simple BERT-based token classification model can perform surprisingly well for end-to-end entity linking, we use a similar neural model for concept wikification: the model takes a sentence (a sequence of words) as input, passes it through the BERT (Devlin et al., 2019) architecture, and then outputs one concept name per token. To train this model for concept wikification, each sentence for training is paired with a sequence of concept names. Therefore, during decoding, the model is going to output a sequence of tags consisting of either concept names or O. An example is illustrated in Figure 2. We refer readers to (Broscheit, 2019) for more details about the model. Since most of the concepts are from scientific literature, we use SciBERT (Beltagy et al., 2019), a BERT variant that is pretrained on 1.14M scientific papers.

Concept type (%)	Example concepts	Example sentences
Disease (16%)	interstitial lung disease, pulmonary hypertension	Other associated lung conditions include **interstitial lung disease**, chronic diffuse **pulmonary hypertension**, pulmonary emboli
Infectious disease (5%)	severe acute respiratory syndrome coronavirus 2	The coronavirus disease 2019 is an ongoing viral pandemic of **severe acute respiratory syndrome coronavirus 2**
Anatomical structure (4.3%)	blood vessel, capillaries	... first to record microscopic observations of ..., spermatozoa, blood and **capillary** flow in **blood vessel**

Table 2: Example concept types, concepts and sentences in the curated COVID-19 concept wikification dataset. The concepts span a large number of types: *disease* (16%), *infectious disease* (5%), *anatomical structure* (4.3%), *academic discipline* (3.6%), *essential medicine* (3.1%), *chemical compound* (3%), *cause of death* (3%), *medication* (2.6%), *symptom* (2%), *medical specialty* (2%).

This hybrid system combines the best of both worlds: it uses the majority-class rule-based approach to resolve unambiguous concept mentions, and a deep contextualized neural model to resolve the ambiguous mentions, which require context to disambiguate.

5 Experiments

We evaluate three wikifiers using the *ambiguous* dataset in Table 1. Evaluating on the *full* dataset does not provide much more insight given that the additional concepts are trivial to resolve with the majority-voting string-matching baseline. The dataset was split into a training set, a development (dev) set and a test set. 60% of the data constitutes the training set, and the remaining 40% is split evenly among dev and test.

We implemented three concept wikifiers:

- **Baseline**: a majority-class baseline simply resolves each mention (a string match of any concept) into its majority-class concept.
- **SciBERT**: a SciBERT-based wikifier, which was trained on the training portion of our curated *ambiguous* dataset. We use frozen weights from SciBERT and only train the token classification layer.
- **SciBERT-FT**: same as SciBERT except that we additionally finetune ("FT") SciBERT.

For pre-processing we use SciBERT's built-in tokenizer to tokenize the sentences. We also use a maximum sentence length of 200 (short sentences will be padded to 200). We train both SciBERT-based models for 50 epochs using a batch size of 200. We use the Adam optimizer with an initial learning rate 5×10^{-5}.

We use micro-average Precision (P), Recall (R) and F1 metrics, for which we only consider a concept link is true if they match the reference annotation. Similar to (Broscheit, 2019), we report two types of P/R/F1 based on whether we require strong match or weak match when calculating these metrics. Strong match requires every token in the gold annotated span to be classified correctly, while weak match only requires at least one token in the gold annotation span to be classified correctly to account for annotation inconsistencies in Wikipedia.

Model	Weak			Strong		
	P	**R**	**F1**	**P**	**R**	**F1**
Baseline	0.32	0.80	0.45	0.32	0.80	0.45
SciBERT	0.72	0.46	0.56	0.59	0.26	0.36
SciBERT-FT	0.67	0.77	0.72	0.66	0.73	0.69

Table 3: Performances of the three concept wikifiers on the ambiguous test dataset. **Strong** and **Weak** are strong match and weak match, respectively. For the majority-class baseline, the strong match scores are the same as the weak match scores.

Experimental results in Table 3 show that the majority-class baseline is insufficient on this dataset, which indicates that this problem is challenging as many concept mentions are ambiguous and refer to more than one concepts. SciBERT-FT achieves significantly higher performance compared to the baseline and the model using frozen SciBERT. This shows that deep contextualized word representations need to be fine-tuned in order to work well for this challenging problem.

6 Conclusion and Future Work

In this paper, we curated a dataset to enable concept wikification research for COVID-19. We developed a few concept wikifier and shown that they perform reasonably well. As a next step, we plan to augment the dataset to include more concepts and will improve the concept wikifier using the augmented dataset.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Yuhang Guo, Bing Qin, Yuqin Li, Ting Liu, and Sheng Li. 2013. Improving candidate generation for entity linking. In *International Conference on Application of Natural Language to Information Systems*, pages 225–236. Springer.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Ying Lin, Chin-Yew Lin, and Heng Ji. 2017. List-only entity linking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 536–541, Vancouver, Canada. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Dan Roth, Heng Ji, Ming-Wei Chang, and Taylor Cassidy. 2014. Wikification and beyond: The challenges of entity and concept grounding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials*, page 7, Baltimore, Maryland, USA. Association for Computational Linguistics.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning distributed representations of texts and entities from knowledge base. *Transactions of the Association for Computational Linguistics*, 5:397–411.

Jin G Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2015. Entity linking for biomedical literature. *BMC medical informatics and decision making*, 15(S1):S4.