A Pipeline for Manual Annotations of Risk Factor Mentions in the COVID-19 Open Research Dataset

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson, Rickard Domeij
The Language Council of Sweden, the Institute for Language and Folklore, Sweden
firstname.lastname@isof.se

Abstract

We here demonstrate how a set of tools that are being maintained and further developed within the Språkbanken Sam and SWE-CLARIN infrastructures can be employed for creating manually labelled training data in a low-resource setting. As example text, we used the "COVID-19 Open Research Dataset", and created manually annotated training data for its associated Kaggle task, "What do we know about COVID-19 risk factors?". We first used our *topic modelling tool* to i) select a text set for manual annotation, ii) classify the texts into preliminary classification categories, and iii) analyse the texts in search for potential refinements of the annotation categories. We then annotated the text set on a more granular level by labelling the token sequences that indicated the existence of the refined categories in the text. Finally, we used the granularly annotated text set as a seed set, and applied our *active learning tool* for actively selecting additional texts for annotation. For the token-sequence annotations, we used our *text annotation tool*, which includes support for incorporating automatic pre-annotations.

1 Introduction

The COVID-19 Open Research Dataset (CORD-19) is a free resource with scholarly articles on viruses from the coronavirus family, and on related topics (Wang et al., 2020). Associated with the dataset is the Kaggle COVID-19 Open Research Dataset Challenge (Allen Institute For AI, 2020), which consists of nine different tasks, all with the aim of extracting from the data what has been published regarding different COVID-19-related research questions.

In order to demonstrate how a set of tools that are being maintained and further developed within the Språkbanken Sam and SWE-CLARIN infrastructures can be combined into a pipeline and employed for creating a manually labelled text corpus, we used the CORD-19 dataset as an example text set. In particular, the aim was to demonstrate how the tools can be useful in a low-resource setting, i.e. with no existing previously annotated data, and with only limited time available for performing manual annotations.

As the example task, for which we aimed to create manually labelled data, we selected one of the Kaggle tasks associated with the dataset, the task "What do we know about COVID-19 risk factors?". For our tool demonstration, we used the version of the CORD-19 dataset that was made available in spring 2020, which contains around 40,000 full text articles.

The first tool used in the pipeline was our topic modelling tool, *Topics2Themes* (Skeppstedt et al., 2018). The tool was used for i) *selecting* data for manual annotation, ii) *classifying* the data into preliminary classification categories, and iii) *analysing* the text material in search for potential refinements of the preliminary annotation categories.

The second tool used was our Språkbanken Sam tool for manual text annotation of token sequences. We employed this tool to manually annotate the texts selected using Topics2Themes according to the refined annotation categories established in the previous step. The annotations were in the form of token sequences that indicated the existence of one of the refined categories in the text.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson and Rickard Domeij 2021. A Pipeline for Manual Annotations of Risk Factor Mentions in the COVID-19 Open Research Dataset. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 214–225.

The third tool employed in the pipeline was our tool for active learning, *PAL*. We used this tool for actively selecting additional texts to manually annotate. These further annotations were also carried out using our Språkbanken Sam tool for manual text annotation. PAL not only actively selects what texts to annotate, but also provides these texts with pre-annotations, using the same machine learning model it uses for text selection. Our manual text annotation tool allows the annotator to correct these pre-annotated labels, as well as to provide the text with new labels. PAL, together with our text annotation tool, thus combines parts of the functionality that is made available by i) a tool such as WebAnno (Yimam et al., 2014), in which annotation suggestions are provided, with ii) the functionality of tools such as JANE (Tomanek et al., 2007) and WordFreak (Morton and LaCivita, 2003), in which data selection through active learning is carried out.

Qualitative text analyses, similar to our approach in the first step of the pipeline, have been conducted in previous research as a preparation for creating annotation categories in the medical domain (Mowery et al., 2012). However, we are not aware of any previous studies that have used a topic modelling tool like Topics2Themes for this task. We are also not aware of previous studies in which this approach has been combined with a pipeline that includes active learning and pre-annotation.

We will here present how we used the three tools in a pipeline for producing annotated texts. The annotation tool will be presented and discussed twice, first when it is used for annotating the texts selected with the topic modelling tool, and thereafter when it is used with pre-annotation of texts selected in the active learning process. We will, however, start by briefly presenting the Kaggle task and our approach for using a set of risk factor keywords for compiling a "Risk factor sub-corpus" from CORD-19.

2 The Kaggle Task and Our "Risk Factor Sub-Corpus"

For the "What do we know about COVID-19 risk factors?"-task, the participants are asked to find out what epidemiological studies report about potential risk factors for COVID-19. Our suggestion for how to approach this task is to train a model to recognise language expressions that are used for describing risk factors for diseases in general, i.e. expressions that we could call "risk factor triggers". To be able to train such a model, a text corpus where such expressions have been manually labelled is needed.

As the first step for creating such a corpus, we compiled a "Risk factor sub-corpus" by extracting 30,000 paragraphs from CORD-19 that contained a risk factor seed word. We used a list of 104 seed words[1] which we had compiled in the following manner: (i) We first constructed a list of the words (or sometimes bi- and tri-grams) that occurred in the call for the "What do we know about COVID-19 risk factors?"-task, and that we estimated would be good seed words for more general text about risk factors. These included, for instance, "risk factors", "factors", "co-infections", "co-morbidities", "high-risk", "pre-existing", and "susceptibility of". (ii) We, thereafter, expanded the list by adding synonyms from the Gavagai living lexicon (Sahlgren et al., 2016). (iii) Words in the list that occurred very often in the CORD-19 corpus were then removed from the seed list as unigrams and specified further with bi-grams. E.g., "factors" was removed and replaced by bi-grams such as "socio-economic factors" and "environmental factors". (iv) We finally read some of the paragraphs containing the seed words, in search for new words to add, and found words such as "more common among" and "more likely".

3 Selecting, Classifying and Analysing the Data with Our Topic Modelling Tool

3.1 Method

As our time available for manual annotation was limited, we decided to focus our effort on text paragraphs with a content typical for the 30,000 paragraphs in the risk factor sub-corpus. With limited annotation resources, we are not likely to be able to catch outliers, or even moderately infrequent content, but we might be able to gather data for training a model that catches typical expressions. Finding topics that represent re-occurring content in a text collection, and creating automatic classes in this content, can be done in an unsupervised fashion, for instance by using topic modelling. For this task, we therefore used the topic modelling tool, Topics2Themes. We used the tool's ability to automatically find synonym

[1] https://www.kaggle.com/mariaskeppstedt/trigger-words

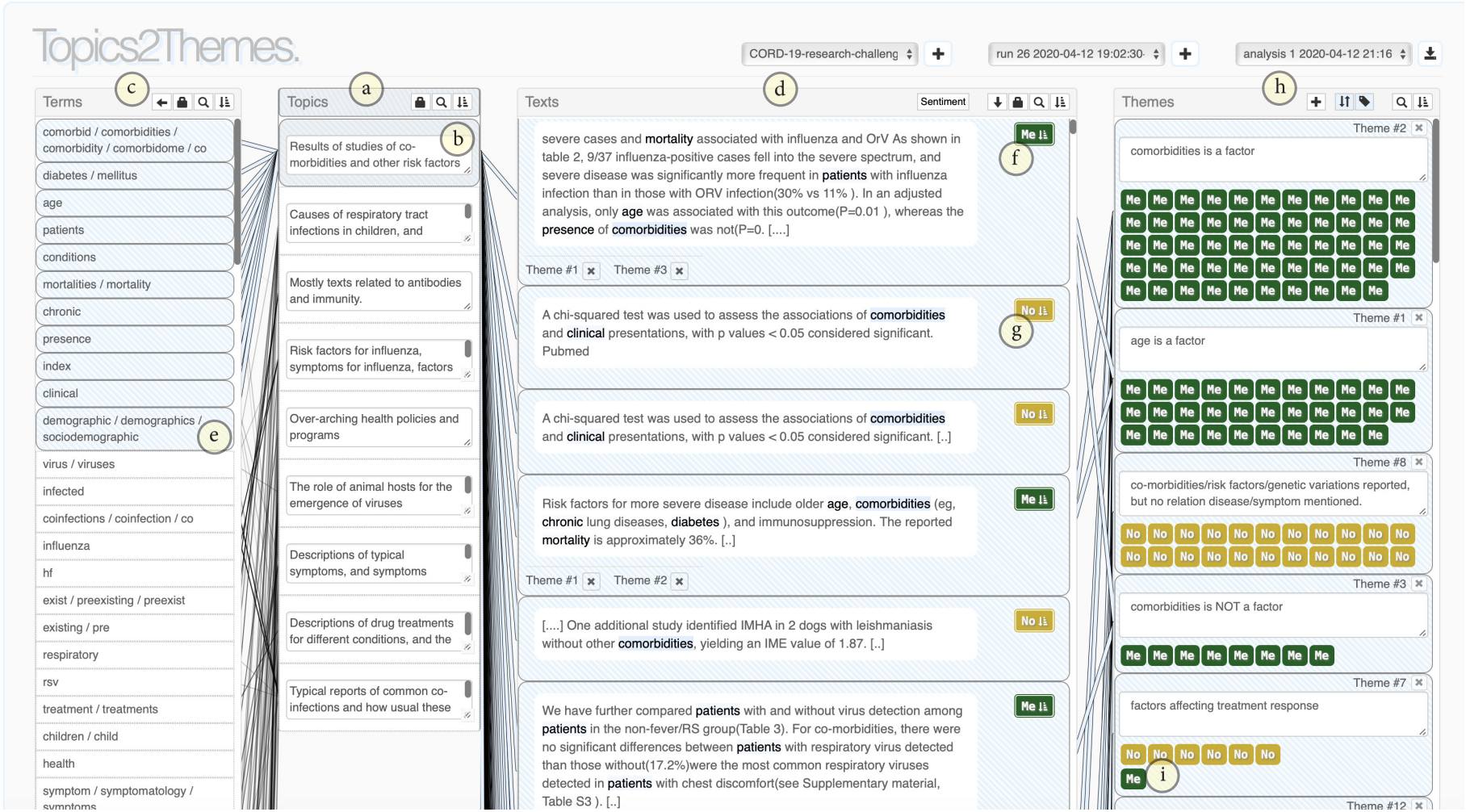

Figure 1. (a) The *Topics* panel contains the nine automatically extracted topics. (b) The user has selected the first topic "Results of studies of co-morbidities and other risk factors" by double-clicking on the topic element in the panel. This has resulted in that the terms associated with this topic, in (c) the *Terms* panel – as well as the texts associated with this topic, in (d) the *Texts* panel – have been sorted as the top-ranked elements in their panels. (e) The Terms panel also shows the results of the automatic synonym clustering, based on word embeddings. (f) Texts with a green label (Me) have been manually classified to contain mentions of risk factors, (g) whereas texts with a yellow label (No) have been classified as not mentioning risk factors. (h) The *Themes* panel contains themes that the user has manually identified when analysing the texts. (i) The theme element labels show how many of their associated texts that have been classified as mentioning risk factors.

clusters in the text with the help of word embeddings. We used embeddings pre-trained on biomedical text (McDonald et al., 2018; Brokos, 2018).

We configured Topics2Themes to use NMF topic modelling to try to find a maximum of 20 topics in the dataset, since our limited annotation resources would not allow us to analyse more than 20 topics. Since NMF is a randomised algorithm, which might produce slightly different results each time it is run, the algorithm was run 50 times, and only topics stable enough to occur in all re-runs were retained. This resulted in nine stable topics being detected, which indicates that it is not likely that the NMF algorithm would have been able to find more than 20 stable topics in this text collection, even if we had configured it to search for more topics. For each topic detected by the topic modelling algorithm, the terms and the texts that are closely associated with the topic are presented by the Topics2Themes tool. For each of the topics extracted, we used the annotation functionality of the tool to manually classify its 15 most closely associated texts as to whether they contained a mention of a risk factor for a disease or not. For the topics that had more than one text among these 15 texts that contained mentions of risk factors, we manually classified additional associated texts in the same manner (up to 70 additional texts for each of these topics).

In addition to classifying the texts for risk factor mentions, the functionality in Topics2Themes for documenting re-occurring themes that are identified when manually analysing the texts was used. The name of the tool, Topics2Themes, reflects this functionality of the tool, i.e., (i) that the automatically created categories – the *topics* detected by the topic modelling algorithm – are presented to the user, and (ii) that the user then manually analyses the output of the algorithm and creates user-defined categories – *themes* – which correspond to the content that the user identifies as important in this output. Typically, such manually identified themes represent re-occurring information in the text collection on a more detailed level than the automatically extracted topics. The aim of the manual analysis was in this case to identify whether there were annotation categories in addition to "describing a disease risk factor" that would be relevant, as well as to determine whether this category should be refined.

3.2 Results and Reflections

Topics2Themes produced nine stable topics for our corpus. Figure 1 shows the user interface of the Topics2Themes tool, and the nine topics detected are shown in its *second* panel. The topic descriptions were given by us, after we had analysed the texts most closely associated with the topics. The figure shows when the topic "Results of studies of co-morbidities and other risk factors" has been selected by the user (as indicated by the blue background). Blue lines connect this selected element with terms associated with the topic (to the left), and texts associated with the topic (to the right).

In the *first* (and left-most) panel, which thus contains the terms associated with the topics extracted, examples of the embedding-based synonym clustering can be seen.

In the *third* panel, which thus contains the texts associated with the topics, each text has an assigned label that is a result from our manual classification. A green label (Me) assigned to the text shows that the text has been classified as mentioning a risk factor for a disease, whereas a yellow label (No) shows that the text has been classified as not containing any information on risk factors. We performed a manual classification for a total of 418 texts, and 150 texts among them were classified as describing risk factors for diseases.

Most of the manually classified texts were associated with five of the topics detected. For these five topics, more than one of the top 15 most closely associated texts described a risk factor for a disease. As described above, we classified additional texts for these topics, in addition to the top 15 most closely associated texts. The five topics were: "Results of studies of co-morbidities and other risk factors", "Causes of respiratory tract infections in children, and whether such previous infections influence the development of asthma", "Mostly texts related to antibodies and immunity", "Risk factors for influenza, symptoms for influenza, factors influencing whether people vaccinate or not", and "Typical reports of common co-infections and how usual these are, and sometimes, also studies of whether they effect severity".

The *fourth* (and right-most) panel contains elements added by the user, in the form or themes identified when analysing the texts. Among the themes identified, 18 re-occurring themes were found. Examples

include "genetics (and family history) is a factor", "co-infection is a factor" and "age is a factor". Five additional examples are shown in the right-most panel in Figure 1. In 24 of the texts analysed, it is described that something could *not* be shown to be a risk factor. Five of the re-occurring themes identified described information of this type. Studies, in which it has *not* been possible to show that something is a risk factor, should be important to identify when mining for risk factors, since the information mined otherwise would be biased towards positive research results. We therefore decided to also include this information in the data to annotate, and consequently classified each such text with the green label that signifies that the text describes risk factors. We also used this output of the analysis – i.e., the frequent occurrences of themes describing that something could *not* be shown to be a risk factor – for refining our preliminary annotation categories, as will be described in the next section[2].

4 Annotation of Token Sequences Using Our Manual Annotation Tool

4.1 Method

We then decided to provide the texts with more granular annotations, which could be useful for training a machine learning model to detect the language that is used for expressing that something is a risk factor for a disease, i.e., what could be called "risk factor triggers". We therefore extended the 150 paragraphs in which disease risk factors were described by labelling the token sequences that the authors had used for expressing that something is a risk factor. That is, we did not annotate the token sequence that describes the disease, nor the token sequence that describes the risk factor, but the language expressions used for indicating that something is a risk factor. For instance the expression "carry a heightened risk of", as shown in Figure 2.

For this more granular annotation, we also – based on the analysis in the previous step – decided to refine the annotation categories by differentiating between if the text described that something was shown to be a risk factor, or if it described that something could *not* be shown to be a risk factor. We consequently created two annotation categories to differentiate between if the token sequences functioned as (i) a risk factor trigger, or (ii) a trigger indicating that something could *not* be shown to be a risk factor. For instance, in the text: *"2019nCoV was of clustering onset, is more likely to infect older men with comorbidities [...]"*, the underlined text was annotated as a risk factor trigger. In contrast, the underlined text in *"The incubation periods did not significantly differ according to age, sex, or the presence of comorbidities [...]"* was annotated as a trigger describing that something could *not* be shown to be a risk factor.

For the token sequence annotations, we used our Språkbanken Sam tool for manual annotation of sequences of text. The user interface of the tool is shown in Figure 2. The tool provides support for two different types of annotations, i) one-token annotations where the annotatable tokens are pre-defined, and ii) annotations of token sequences that are to be IOB-coded. We here used the tool setting that provides support for the latter type.[3] The tool is optimised for annotation speed, both when adding new labels and when changing existing labels. The annotator can either use the mouse to select a single token or to select a sequence of tokens. This results in a pop-up (shown in Figure 2) which allows the annotator to quickly choose which label to use for the token(s) selected. If a B-tag is chosen for a selected sequence of tokens, the first token in this sequence will be given the B-tag, and the subsequent tokens in the selected region will be given I-tags.

4.2 Results and Reflections

The total size of the text set annotated was around 50,000 tokens. In this set, there were a total of 282 token sequences indicating that something is a risk factor, and 44 token sequences indicating that something could *not* be shown to be a risk factor. The types of expressions that we target thus occur very rarely in our corpus, despite the fact that texts with a higher probability of containing mentions of risk factors had been selected for the text set. That is, the 50,000-token text set was not a randomly selected

[2]The full configuration and analysis is available at: https://www.kaggle.com/mariaskeppstedt/cord19clarin2020analysis

[3]The IOB format for the entities in the examples sentence were thus coded as:
[... (is, **B-RISK**) (more, **I**) (likely, **I**) (to, **I**) ...**]** and **[**... (did, **B-NO**) (not, **I**) (significantly, **I**) (differ, **I**) (according, **I**) (to, **I**) ...**]**

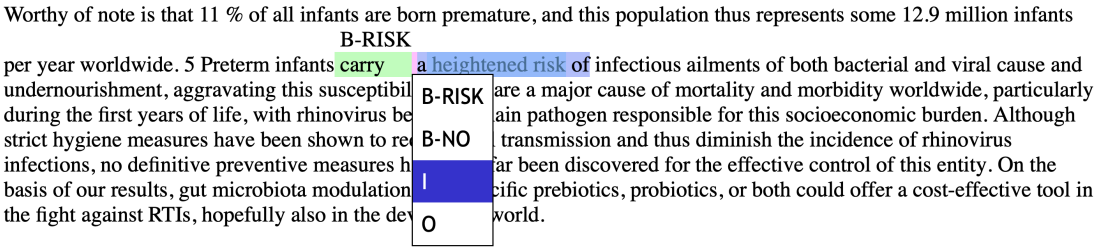

Figure 2. The pre-annotation functionality has labelled the token "*carry*" with the category *B-RISK*. The user has then selected the subsequent four tokens for annotation. This has resulted in a pop-up with the four possible annotation categories: (i) *B-RISK* is the first token – in a sequence of tokens – in an expression which indicates that something is a risk factor, (ii) correspondingly, *B-NO* is the first token in an expression indicating that something could not be shown to be a risk factor, (iii) *I* is the subsequent tokens in the expression (i.e., the category of the first token in an expression determines its type in the annotation tool), and finally (iv) *O* signifies that a token is not included in an expression (this is also the default category for tokens not annotated). The user has here chosen the *I* category, which has the effect that the sequence of tokens "carry a heightened risk of" will be labelled as belonging to the *RISK* category when the annotated data is exported from the tool.

text set, but one with texts that had first been selected based on that they contained keywords for risk factors, and many of them were associated with one of the five topic-modelling-topics that were related to risk factors. Thereby, in our low-resource setting, it would probably not have been fruitful to select a random set of texts for annotation, since it is likely that very few risk factor mentions would have been found.

5 Selecting More Texts for Token Sequence Annotation, Using Our Active Learning Tool

5.1 Method

The third component in our pipeline for creating annotated training data while still only employing limited manual annotation resources was to use our active learning tool, PAL (Skeppstedt et al., 2016). Active learning is a machine learning/data selection technique, where data for manual annotation is actively selected by a machine learning model. Thereby, the machine learning model has the possibility to select the data points – from a large pool of unannotated data – which are most useful for improving the model. For instance, the machine learning model could use uncertainty sampling (Schein and Ungar, 2007; Settles, 2009), i.e., select those data points for which the model is most uncertain regarding how the data should be classified. In a successful active learning set-up, a machine learning model trained on relatively few data points would yield the same performance as a model trained on a larger dataset. It is thereby possible to limit the number of training data points that need to be manually labelled. We have previously conducted experiments with PAL, evaluating the active learning functionality through simulations on labelled datasets (Skeppstedt et al., 2019). For the task of training a model to recognise three different kinds of named entities in tweets, active learning was shown to be more efficient in the use of manually labelled training data than a random selection of manually annotated tweets. That is, models trained on actively selected tweets performed better than models trained on the same amount of randomly selected tweets.

PAL is targeted towards small training datasets, and therefore uses an active learning approach that is more likely to function on small datasets, in the form of uncertainty sampling using a token-level logistic regression classifier. The tool can incorporate features in the form of word embedding vectors

when training the logistic regression classifier. When previously evaluating PAL's performance for named entity recognition in tweets, the incorporation of embeddings was useful for two of the three named entity categories evaluated.

We constructed a pool of 5,000 unlabelled data points, in the form of 5,000 paragraphs from the CORD-19 dataset that had not yet been annotated. That is, we derived the paragraphs for the unlabelled pool from the entire CORD-19 corpus, and not only from our keyword-based "Risk factor sub-corpus". We then used the annotated corpus constructed in the previous steps as the seed set, in order to let PAL train a model to use for active selection of paragraphs from the pool of unlabelled data. We configured the tool to select the 35 most uncertain data points in each active selection/annotation round, but to prioritise texts which the model predicted to contain at least one pre-labelled token. As features, the model was configured to use a concatenated vector consisting of the one-hot encoding of the token to be classified, the one-hot encodings for its four preceding and four following tokens, as well as the embedding vectors for all these nine tokens. We used the same biomedical embeddings (McDonald et al., 2018; Brokos, 2018) that we used for the Topics2Themes tool.

We ran the active learning process in nine iterations. For each iteration, 35 new paragraphs (those containing the 35 tokens for which the machine learning model was most uncertain) were actively selected for manual annotation. After having annotated these 35 paragraphs, they were added as new data samples to the training dataset, i.e., to be used for training the machine learning model for the next active learning iteration.

For each iteration in the active learning process, PAL generates a plot, in order to provide the user with an understanding of how the uncertainty estimations for the unlabelled data pool changes during the process.

5.2 Results and Reflections

Plots generated by PAL for two of the nine iterations are shown in Figure 3. The plots shown are those generated when i) the active learning process is first run (with 418 labelled samples available for training the machine learning model), and ii) the training dataset contains a total of 698 manually labelled samples. The plot included here is shown from the point of view of a model which detects expressions that indicate that something is a risk factor. Thereby, the colour red is used for tokens that the model classifies as risk factor indicators, and blue is used for all other tokens. Corresponding plots (not included here) are generated from the point of view of a model which detects expressions that indicate that something could *not* be shown to be a risk factor.

To the left in the plot, the content of the pool of unlabelled data is visualised, through a t-SNE plot (van der Maaten and Hinton, 2008) of the most frequently occurring words in the data pool. Again, the biomedical embeddings (McDonald et al., 2018; Brokos, 2018) were used. The plot thus shows the semantic distribution of the words in the corpus, where semantically similar words are represented by dots that are positioned close to each other in the plot. The hue of the dot is determined by the token instance of the word for which the model is most uncertain, i.e., the larger the uncertainty for the classification of this token, the darker is the colour with which it is displayed.

To the right of the plot, the 35 tokens for which the model is most uncertain are shown, i.e., the tokens on which the decision for which 35 paragraphs to select for manual annotation is based. The actual token is shown in the center, and to its left and right, its textual context is shown. The bars show the level of uncertainty with which the model has classified the tokens, and the bar colour is determined by the class of the token (as classified by the model). When the model was too uncertain to be able to make a decision for how to classify the token, the bar is shown in black.

Despite the few iterations in which the active learning process was run, the plots generated by PAL show how the state of the pool of unlabelled data changes. After nine iterations, there seems to be less uncertainty left in the data pool. This is most evident by the lengths and colours of the bars representing the tokens, but there is also a small indication through lighter colours in the t-SNE plot and through the bar representing the mean uncertainty in the pool of unlabelled data.

The aim of providing visualisations for the uncertainty left in the pool of unlabelled data is to make the

Expressions indicating that something is a risk factor

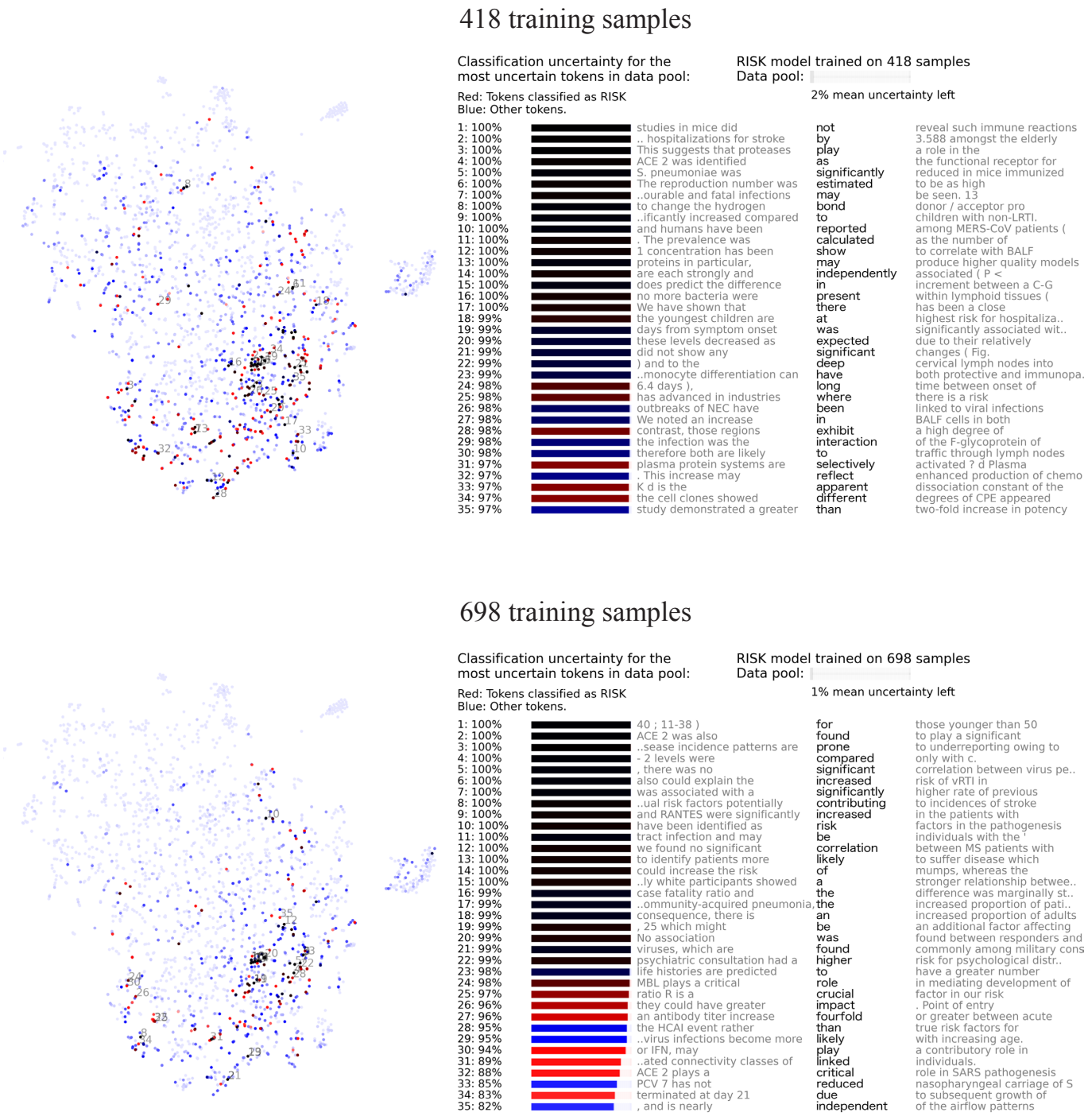

Figure 3. Two of the plots generated by PAL during the active learning process: (i) When the active learning process is first run (with 418 annotated training samples), and (ii) after a total of 698 samples have been annotated and added to the training data. To the left, the total uncertainty in the pool of unlabelled data is shown through a t-SNE plot (the darker the colours of the dots, the more uncertainty is left). To the right, the classifier uncertainty for the 35 most uncertain tokens are shown, i.e., the tokens on which the choice of which paragraphs to select for annotation was based.

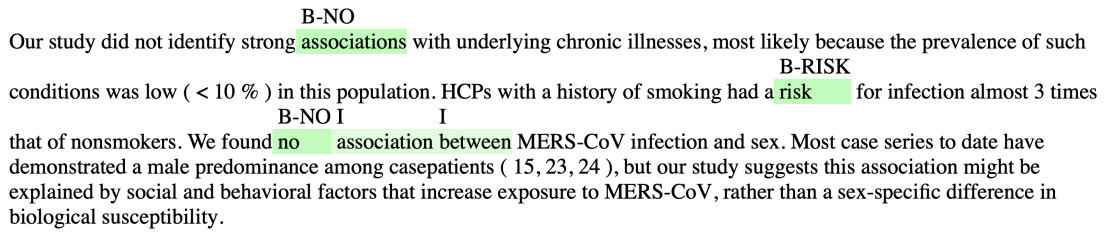

Figure 4. Automatic pre-annotations produced by PAL and imported into the Språkbanken Sam tool for manual annotation.

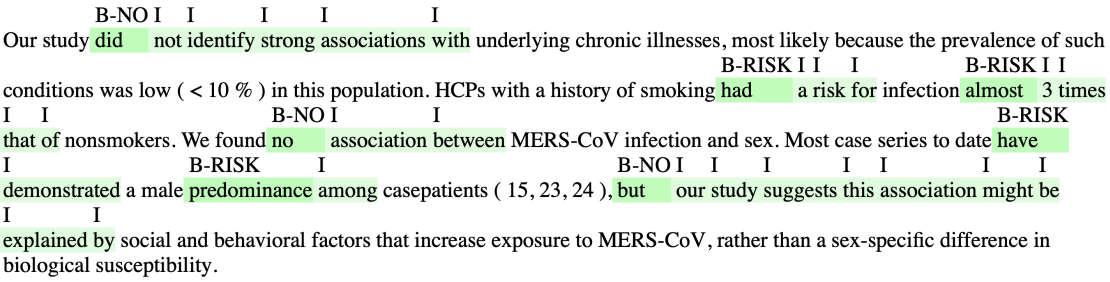

Figure 5. Manual annotations carried out in the Språkbanken Sam annotation tool.

active learning and annotation process more interesting. Thereby, there is a possibility to also increase the annotator's intrinsic motivation for the annotation task. That there was a change in visualised uncertainty levels already after nine iterations shows that there is a potential for using these kinds of visualisations for increasing the interest in the annotation task, also very early in the active learning process.

6 Manual Annotation of Token Sequences with Pre-Annotation

6.1 Method

The same logistic regression model, which is used for actively selecting training data samples, is also used by PAL for providing the selected samples with pre-annotated labels.

The pre-annotations from PAL were in previous versions of the tool only provided in the format of the annotation tool BRAT (Stenetorp et al., 2012). However, BRAT provides a rather extensive set of functions for text annotation, which also has the effect that the procedure for annotating token sequences is more time consuming than when using an annotation tool specifically adapted for this task, e.g., the annotation tool which we have developed at Språkbanken Sam. We have therefore made it possible to also import pre-annotations from PAL into the Språkbanken Sam annotation tool.

Figure 4 shows an example of a pre-annotated text, and Figure 5 shows how manual annotations have been provided to the same text.

6.2 Results and Reflections

Also this second part of the manually annotated corpus contains around 50,000 tokens. For this text set, we found a total of 224 token sequences indicating that something is a risk factor, and 39 token sequences describing that something could *not* be shown to be a risk factor. That is, slightly fewer token sequences were detected in the actively selected sub-corpus, than in the one compiled through topic modelling.

While annotating the texts, we could observe that the quality of the pre-annotations was not very high. This is exemplified by the text paragraph in figures 4 and 5. In fact, this paragraph is chosen as an example paragraph here because it contained many instances of annotated token sequences, not because it was a

paragraph representative for the performance of the pre-annotation. It contains one pre-annotated token sequence that was not at all altered by the annotator, which was quite rare.

We had expected that low-quality pre-annotations would disturb the flow of the manual annotations, and therefore not be perceived as useful by the annotator. However, the opposite was experienced. That is, many consecutive sentences without pre-annotated content were subjectively perceived as boring to annotate, while sentences with one or several pre-annotations were found interesting. That the lower-quality pre-annotations were not found disturbing might be explained by the fact that the annotation tool is optimised for annotation speed, also for altering pre-annotated content. When pre-annotated tokens are selected by the annotator and given a different annotation category than the one provided by the pre-annotations, the pre-annotated content is automatically removed by the annotation tool. It is also easy to change the annotation category of a sequence, but to keep the annotated token span, by just changing the annotation category of the first token. The simplicity with which pre-annotations can be altered comes with a trade-off, since – unlike for the BRAT tool – a token cannot be assigned to several categories with this set-up. However, with an annotation task that only allows one category per token, we believe it is better to choose a tool that is optimised for annotation speed.

The sentiment towards pre-annotations with a lower quality that we present here was a subjective assessment made by the annotator in this study. The attitude towards lower-quality pre-annotations might vary between annotators, and not everyone might find that the presence of pre-annotations makes the annotation task more interesting. However, one of the lessons learnt here is that it might be worth to at least give the annotator a choice to include pre-annotations, also pre-annotations with a lower quality. That said, pre-annotations with a very low quality are probably not found useful by any annotator.

7 Discussion and Tool/Data Availability

Our resources for manual annotation were scarce, not only in terms of the number of man-hours that we were able to spend on creating the annotated corpus, but also in terms of competence within the medical domain. One of the authors had previous experience in annotation guideline creation and medical text annotation in collaboration with physicians, and was also the one who carried out the manual annotations. However, without a medical education, some text content is difficult to understand, e.g., to distinguish between risk factors for a disease, and causes, signs and symptoms associated with the disease. Even more difficult is the development of comprehensive annotation guidelines, without access to medical knowledge, e.g., guidelines regarding which annotation categories to include and regarding exactly what should be counted as a risk factor for a disease. We, therefore, did not construct any detailed annotation guidelines, apart from the short description of the two annotation categories given above.

While the fact that we lack medical competence might decrease the value of the annotated corpus created, it also highlights the importance of annotation pipelines, similar to the one we have demonstrated here. That is, annotation pipelines with the potential of supporting annotation guideline creation, facilitating annotation, and minimising the amount of annotated data required. While it is possible to obtain laymen annotations for English texts at a low cost, annotations carried out by annotators with extensive medical knowledge tend to be more expensive. Thereby, it is important to use the resource of medical expertise wisely. A topic modelling tool might give the medical expert an overview of typical categories in the texts, which might help in determining annotation categories and creating guidelines. An annotation tool with a high usability, and through which the annotator is able to track the status of the active learning process, might make it faster to annotate and increase the expert's intrinsic motivation for the annotation task. Finally, an active selection of training data samples that are useful for a machine learning model might make it possible to train useful models without having to manually annotate very large corpora.

Both Topics2Themes[4] and PAL[5] are freely available on GitHub. We plan to continue the development of our tool for manual text annotation, and to also make this tool freely available. We will also continue the development of Topics2Themes. After this study was conducted, we have added the functionality of allowing the user to provide a manually constructed list of multiword expressions, i.e., expressions that

[4]https://github.com/mariask2/topics2themes

[5]https://github.com/mariask2/PAL-A-tool-for-Pre-annotation-and-Active-Learning

are then treated as any other word by the topic modelling algorithm. This functionality could, however, be extended by also providing an automatic detection of multiword expressions.

We have also made our two annotated datasets, i.e., the set selected through topic modelling and the set selected through active learning, freely available at Kaggle.[6] The two datasets consist of around 100,000 tokens, with a total of 506 token sequences annotated as expressions used for describing that something is a risk factor, and 83 token sequences annotated for descriptions of when something could *not* be shown to be a risk factor. Although the small size of these annotated datasets might not be sufficient for training high-performing classification models, we welcome anyone to use these annotations for classifier experiments, or to use them as seed sets in active learning and/or pre-annotation approaches for further expanding the training dataset. We would also appreciate efforts from others – in particular annotators with a medical background – to annotate the same dataset, to be able to compute inter-annotator agreement, or to use the annotations as a support for developing a set of detailed annotation guidelines.

Our next step will consist of making use of data contributed by others at Kaggle. For the risk factor task, there is structured data collected regarding studies of COVID-19 risk factors, together with links to relevant articles. These articles might be used for collecting and annotating a text set that can be employed as an independent gold standard, against which our approach for creating a training dataset for risk factor mentions can be evaluated.

Although the main purpose of this study has been to demonstrate the use of different types of tools for the creation of an annotated dataset – rather than the resulting dataset – we still consider this type of data, and its associated Kaggle task, as important. Natural language processing tools that can help researchers to access the content of scientific papers regarding risk factors for COVID-19 are useful, for instance when criteria for COVID-19 vaccine prioritisation must be established. Such tools can be developed using the kinds of annotated datasets that we have created here, and methods for efficiently creating these datasets are therefore important.

Acknowledgements

This work was supported by the Swedish Research Council (2017-00626).

References

Allen Institute For AI. 2020. COVID-19 open research dataset challenge (CORD-19). https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks.

Georgios-Ioannis Brokos. 2018. Biomedical pre-trained word embeddings. https://github.com/RaRe-Technologies/gensim-data/issues/28.

Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas Morton and Jeremy LaCivita. 2003. Wordfreak: An open tool for linguistic annotation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations - Volume 4 (NAACL HLT)*, pages 17–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Danielle L Mowery, Sumithra Velupillai, and Wendy W Chapman. 2012. Medical diagnosis lost in translation – analysis of uncertainty and negation expressions in English and Swedish clinical texts. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 56–64, Montréal, Canada, June. Association for Computational Linguistics.

Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan, and Anders Holst. 2016. The Gavagai living lexicon. In *Proceedings of the Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Andrew I. Schein and Lyle H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Mach. Learn.*, 68(3):235–265, October.

[6] https://www.kaggle.com/mariaskeppstedt/manually-annotated-risk-factor-expressions

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report #1648, University of Wisconsin–Madison, http://research.cs.wisc.edu/techreports/2009/TR1648.pdf.

Maria Skeppstedt, Carita Paradis, and Andreas Kerren. 2016. PAL, a tool for Pre-annotation and Active Learning. *JLCL*, 31(1):91–110.

Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018. Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 9–16.

Maria Skeppstedt, Rafal Rzepka, Kenji Araki, and Andreas Kerren. 2019. Visualising and evaluating the effects of combining active learning with word embedding features. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 91–100. German Society for Computational Linguistics and Language Technology (GSCL).

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Efficient annotation with the Jena ANnotation Environment (JANE). In *Proceedings of the Linguistic Annotation Workshop*, pages 9–16, Stroudsburg, PA, USA, June. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Muhie Seid Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Stroudsburg, PA, USA. Association for Computational Linguistics.