Novel Disease Treatment Identification Through Deep Learning From Podcasts

Abstract

Identifying previously unknown treatments for diseases is an important problem in medicine. We identify podcasts as a potential source of knowledge from which we can extract possible novel treatments for various diseases. Several podcasts from People's Pharmacy were taken and transcribed using Google Cloud services; this document outlines a possible method for extracting novel treatments for diseases from these transcripts.

Data

We develop the proposed method below using some excerpts from the data, using a few excerpts to demonstrate the ideas and reasoning behind the method. We use bold typeface to denote medical terms, italics to denote words to also look out for, and underlined to denote medical terms we do not care about.

Excerpt 1

> ... in the People's Pharmacy Health headlines new *research confirms* your grandmother was right getting enough **sleep** keeps your **immune system** in good working order the **study** included 164 healthy young men and women who kept track of their **sleep** every night for a week using Diaries and devices than they were deliberately exposed to a standard quantity of **cold virus** which was sprayed into their noses after the exposure the volunteers were quarantined in monitor to see if they started **sniffling sneezing** and **coughing** within 5 days 45% of those who averaged less than 5 hours of **sleep** nightly came down with **cold symptoms** compared to just 17% of those who slept at least 7 hours most nights the author's point out that previous **research** has shown **immune** system impairment in people who are sleep-deprived they did not study **immune** system activity in these volunteers nor does anyone know whether interventions to improve sleep such as medication ...

- We could possibly train a one-one LSTM on the bold terms. Since the underlined terms also may be chosen, we need to set a threshold. Alternatively, we could train a language model, and see if the medical terms can be clustered together. If so, we find a suitable radius, and map each word in the text to that vector space and see if it falls in that n-ball, and ignore words that fall outside the n-ball.
- We could deal with the underlined terms by making a hand-crafted list of terms to ignore and removing them before passing the text to the model.

Excerpt 2

> ... nor does anyone know whether interventions to improve **sleep** such as medication for **cognitive behavioral therapy** would boost **immune** defenses against **infection** serious complication of certain popular **diabetes** drugs **gliptins dpp-4 Inhibitors** include medication Januvia and onglyza by the FDA in 2006 and how to reach the pharmacy shelves in 2009 now after several years the agency has determined that these and similar *drugs* can cause severe and potentially disabling **pain** the **pain** May begin within a few days of starting the medicine but some people report that it can take years to show

> up in most cases it goes away within a month of stopping the medication the FDA encourages people to report to their prescribers promptly ...

- Note that we may have duplicate words, like "pain" here. It is a design decision to choose whether to keep them or collapse them.
- After we've run the text through the RNN in the test phase, we end up with a vector of 1s and 0s. Alternatively, we could do multi-class classification where 0 = ignore, 1 = drug/agent; 2 = target/symptom/disease). If we use binary classification, we could
    - average out the values and set a threshold (may not work, may cause false negatives).
    - find pairs and then check with the existing graph, perhaps using a BFS-like algorithm. This would be easier with the multi-class classification.

Excerpt 3

> ... on February 22nd 1997 Dr Oliver Sacks came into our studio with a wet bathing suit in one hand and a cycad plant in the other he was there to talk about his book The Island of the color-blind doctor sex died on August 30th to honor his memory we shared this conversation about one of his lesser-known books how does the **environment** affect **neurological** function coming up on the People's Pharmacy a rebroadcast of our 1997 interview with dr. Oliver Sacks ...

- This particular extract had a confidence of 0.969. Note however, that there are typos, e.g., "doctor sex" instead of, "Dr. Sacks". It may be a hindrance if such typos also existed across medical terms. We may have to do an estimation of the percentage of such typos in our data.
- Of course, if we used a threshold, this particular extract wouldn't even be selected, but it's easy to see how typos could be present in more important extracts as well.

> ... persistent **joint pain** sometimes leads people to sign up for a **hip** or **knee replacement** a new *study* of around 20,000 people who went through **replacement surgery** has found that the chance of a **heart attack** is higher in the month following the **surgery** those who had a **knee replacement** at eight times the heart attack risk in that first post-surgical month while for those whose hips were replaced the risk was four times higher The increased risk of a **heart attack** did not last much beyond the first month after the surgery but a risk of **blood clots** lasted for several years after the surgery was done these can also be extremely dangerous if they break loose and launched in the **lungs** or the **brain** low levels of **vitamin D** May contribute to the chance of developing age-related **macular degeneration** but only in people with genetic susceptibility ...

- Note the first sentence saying, "X *sometimes leads people to do* Y". We should try and ignore the entire sentence with such phrases. But without periods, we need a way to identify the end of that sentence.
- This paragraph indicates that knee replacement surgery increases the risk for, and therefore may be a cause of, heart attacks. Multi-class classification followed by pairwise matching should, in theory, catch this.

> ... scientists at the University of Buffalo analyzed data on more than 1,200 women in a section of the Women's Health Initiative this part of the *study* is considering **nutritional** status and the risk of **macular degeneration serum** levels were assessed as **25-hydroxy Vitamin D** which is a marker of **vitamin D** from all sources the likelihood of **macular degeneration** was strongest in women with a specific variant of an **immune** system **Gene** called **cfh** who also had low levels of **25-hydroxy vitamin D** ...

- Note again the duplicate vitamin D and macular degeneration terms. Such duplicates could give the extract a false sense of importance.

Proposed Method

Therefore, the method has the following steps:

1. **Low confidence text pruning:** Remove excerpts with confidence below some threshold, say 96%.
2. **Irrelevant term removal:** Remove medical terms that aren't useful, such as "Pharmacy", "medicine", or "prescribe", and all forms of these words from the excerpts.
3. **Classification:** Run an LSTM trained to classify words as important or not important on the excerpts. This is a form of named entity recognition, but training may involve manually labeling several excerpts. Alternatively, a multi-class version of this can be performed, where important terms are further divided into "agents" such as drugs or treatments, and "targets" such as diseases.
4. **Speculation removal:** From each excerpt, remove phrases of the form, "X leads people to do Y". Because the data does not have periods, we could find such speculative phrases and ignore that specific pair of important words around them.
5. **Adjacent duplicates removal:** Remove duplicates that are adjacent to each other among important words in each excerpt.
6. **Pair formation:** Retrieve adjacent pairs of important words. If multi-class classification was used, then only retrieve pairs of agents and targets.
7. **Validation:** Check with a known database of pairs of agents and targets.