Progress on Extraction of Medical Guidelines

Sean Gallagher, Hossein Hemati, Wlodek Zadrozny

March 15, 2015

Abstract

Our objective in the project has thus far been to show that we can extract medical information from text, particularly medical guidelines from text. In doing this, we have taken three approaches, building links from analysis of constituent trees, building links from regular expression matches on plain text, and via answer type detection in an existing question answering system.

1 Constituent Trees

Our first attempt at extracting medical conditions and symptoms from unstructured text is by using heirarchical analysis of a constituent parse tree of the text. The system is relatively simple, following a general pattern:

1. Decend to the leaves of the tree, the original words.
2. Match classes of words for operators and numeric values.
3. Match the remaining words against known tests and diseases.
4. Ascend the tree, combining analyses as follows:
    - A decimal value, an operator, and a test constitute a condition.
    - A condition and an ailment constitute a link, which is the program output.

There are several concerns about this approach, not the least of which are the limited scope of the links it can extract. It expects guidelines of the form "Values of $A1C \geq 6.5$ may indicate diabetes," but this causes serious brittleness. Consider the generally equivalent statement: "A1C values predict diabetes when greater than approximately 6.5."

There are also weaknesses with tables, which are relatively common in medical literature, and any document of such links. Take for example the Wikipedia article on Diabetes. The condition information is stored in a structured way, as a table. The trouble is that it is not clear how to interpret the contents. "≥ 6.5" can be interpreted as a operator and value. But the rest of the content needs more context, and this is what we are aiming to solve in section three.

WHO diabetes diagnostic criteria[40][41] edit

Condition	2 hour glucose	Fasting glucose	HbA_{1c}	
Unit	mmol/l(mg/dl)	mmol/l(mg/dl)	mmol/mol	DCCT %
Normal	<7.8 (<140)	<6.1 (<110)	<42	<6.0
Impaired fasting glycaemia	<7.8 (<140)	≥6.1(≥110) & <7.0(<126)	42-46	6.0–6.4
Impaired glucose tolerance	≥7.8 (≥140)	<7.0 (<126)	42-46	6.0–6.4
Diabetes mellitus	≥11.1 (≥200)	≥7.0 (≥126)	≥48	≥6.5

Figure 1: Excerpt of the Wikipedia article on Diabetes

2 Regular Expressions

3 Dependency Graphs

We aim to improve our recognition of diagnostic criteria using other knowledge also available on Wikipedia. We can do that with many of the tools we have already been developing for question answering with Watsonsim.

Firstly, rather than hard-coding the names of many diseases we want to detect, we can recognize the types of the words we encounter using answer type detection. This way our approach will generalize better for new diseases. Secondly, we can use many Wikipedia sources to determine semantic relations between phrases. For example, we have determined that HbA_{1c}, as seen in the excerpt given above, is a synonym for the main article "Glycated hemoglobin." We do this by keeping a table of the redirects between Wikipedia articles. We also assume that phrases to label a link to an article are synonymous with the title of the target article, and we use this to measure synonymy as well.

3.1 Approaches

On the outset we intended to use DBPedia to detect answer types, but these are not usually specific enough to be greatly helpful, and are synonymous with the target type only on very rare occasions. We attempted to supplement this using ConceptNet, but our initial tests, before parsing the entire knowledgebase, still do not indicate very high recall rates. Instead, we have developed a simple rule-based method of extracting lexical answer types from supporting passages, and we are examining whether to include the NELL ontology, which does at least contain many of the phrases we are targeting.

Unfortunately, this is not yet become helpful because there is no reason to believe from Wikipedia's synopsis that Glycated hemoglobin is a medical test. Instead, it states:

> Glycated hemoglobin (hemoglobin A1c, HbA1c, A1C, or Hb1c; sometimes also HbA1c or HGBA1C) is a form of hemoglobin that is measured primarily to identify the average plasma glucose concentration over prolonged periods of time.

3.2 LAT Detection

3.2.1 Next Steps

3.3 Link Expansion