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.

```
(ROOT
  (NP
    (NP (NNP Executive) (NNP Summary))
    (: :)
    (NP
      (NP (NNPS Standards))
      (PP (IN of)
        (NP (NNP Medical) (NNP Care)))
      (PP (IN in)
        (NP (NNP Diabetesd2013) (NNP Current) (NNS criteria)))
      (PP (IN for)
        (NP
          (NP (DT the) (NN diagnosis))
          (PP (IN of)
            (NP
              (NP (NN diabetes) (CD •) (NN A1C))
              (JJR >) (JJ =) (CD 6.5) (NN %))))))
    (. .)))
```

Figure 1: Constituent Parse of a Matched Sentence

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 2: Excerpt of the Wikipedia article on Diabetes Mellitus

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 which label a link to an article are synonymous with the title of the target article, and we use this to measure synonymy.

3.1 Lexical Answer Types

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 knowledge base, 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. This covers many cases and are reasonably specific but in several instances they make unnecessary distinctions. (Sign and symptom may often be used interchangably, even though Wikipedia makes a distinction; we have noted similar situations with writer and author.) At any rate, we are examining whether to include the NELL ontology, which does at least contain many of the phrases we are targeting.

Nonetheless, there are situations where knowing the exact type of a referent is not immediately helpful. Going back to the previous example, there is no direct statement that glycated hemoglobin is a medical test in the following passage, because it is actually the subject of an implied test of its quantity:

> 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.

This is not immediately helpful because it is not yet clear that glycated hemoglobin is the subject of any test. Unless NELL or a similar ontology has this information, some other general solution for reading such tables will need to be found, or else they may need to be handled more manually.

3.2 Link Expansion

When evaluating supporting passages to discover answer types, we have found it useful to annotate more than just the answer the passage supports. Finding a new object with the

correct LAT but in a supporting passage for another candidate answer now triggers the generation of new candidate answers.

The intuition for such an inclusion is clear. The symptom of pain is relevant enough to diabetes for authors to choose to include it, but diabetes is not relevant enough to pain for the inverse to occur. As a result, any list of symptoms or conditions relating to diabetes will be incomplete unless the evidence for an answer comes in large part from outside of its article.

The new implications of the LAT search has the curious effect that the text analysis pipeline operates as a tree when a detected LAT suggests a new candidate. We believe we can further stimulate this effect by then running LAT detection on the newly generated candidate answers, making a pipeline a graph, but the appropriate termination conditions are as yet unclear.

3.3 Examples and Progress

The dependency graph approach is not complete enough to extract full conditions, though it can find symptoms and signs. For example, a query for the symptoms of diabetes will find the source text that includes the following ten symptoms for diabetes:

frequent urination	increased hunger	blurry vision	diabetic dermadromes	fatigue
increased thirst	weight loss	itchy skin	slow healing of cuts	headache

- The query, which is formed to imitate Jeopardy (although this soon will not be necessary), is the following: "This symptom indicates diabetes:" The suggestion apparatus adds muscle pain and thirst as symptoms, but only thirst makes the list of top ten candidate answers.
- For the closely related query "This sign indicates diabetes," "Proteinuria" is the top result, although in this case it is read from the wiki page on the symptom rather than the disease.
- The query "This test indicates diabetes" discovers the type of the A1C test, and also suggests the answer "blood glucose fasting test" but later scoring does not rank the candidate very highly.

As you can see, the current recall of the LAT method is not yet more than about 10%. The precision of answers which are actually symptoms is high but the still fewer than one in ten candidates are symptoms. This reason becomes rather clear in resulting candidate answer list of the last query. Answers of the type requested are buried among answers which only mention the type.

```
0: [0.862925 blllllllll] Diabetes Tests and diagnosis
1: [0.837302 bbllllllllll] Glucose Tests
2: [0.755305 blllllllll] Testing
3: [0.686052 blllllllll] Blood Tests that Indicate Diabetes
4: [0.644805 blllllllll] The A1C Test and Diabetes
```

```
5: [0.644578 blllllllll] Type 2 diabetes Tests and diagnosis
6: [0.572282 blllllllll] Diagnosis of Diabetes and Prediabetes
7: [0.572118 blllllllll] Tests to Diagnose Diabetes
8: [0.559373 blllllllll] Type 2 Diabetes
9: [0.554920 blllllllll] 5 Important Tests for Type 2 Diabetes
```

This could be seen as a failure of scoring, but to be fair to the model, the LAT feature has rather low recall and hence few instances to extract a coherent meaning from. This is something that needs to be addressed, and is currently the subject of development. Much low hanging fruit remains, including existing knowledge bases, more complete rules for extracting type-implying syntax, and better rules for matching hypo- and hypernyms among types.