SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1
Leading health innovation through data

201911 Generating SNOMED CT-encoded test data vignettes using Wikipedia

Edward Cheetham, NHS Digital (United Kingdom)

Summary

This presentation demonstrates the way a structured knowledge resource (Wikipedia), SNOMED CT and a simple NLP pipeline can be combined to generate simple clinical vignettes suitable for use as test data. Lessons learned, opportunities for product improvement and further work are discussed.

Audience

Research/academic, Technical , Clinical

Learning Objectives

1. Demonstrate the ease of using SNOMED CT in an NLP pipeline for analysis of clinical knowledge resources
2. Demonstrate examples of test data that can be produced
3. Consider the opportunities for clinical engagement and content development that this approach may allow

Abstract

Introduction

Good quality clinical test data is valuable for many health information technology development purposes. However, test data is difficult to produce, either by anonymising real clinical data (with confidentiality concerns) or by generating suitable synthetic data (which can be time-consuming). Wikipedia is an online encyclopedia comprising more than 40 million articles on a wide range of topics, and has many entries describing human diseases, their features and management. Wikipedia content can be accessed several ways, including via an API. This paper reports on using Wikipedia as an example of a structured clinical knowledge resource, analysing the text of sample entries using a simple natural language processing (NLP) pipeline with SNOMED CT as a structured entity recognition resource, and generating coded clinical vignettes for use as test data.

Methods

Wikipedia page detection

This project used the MediaWiki API [1] to automate the identification of Wikipedia entries that correspond to SNOMED CT-encoded entries. Membership of the GP/FP RefSet were used to produce a ‘search list’. Terms for each concept were passed to the API search function, returning the plain text content of matched page for further testing. Currently bulleted list and tabular data is often not returned with this approach.

Searches can match the submitted term exactly or can make use of Wikipedia’s auto-suggest feature. Generally, the auto-suggest feature increases sensitivity. Very occasionally matches are only identified with exact matching, so using both approaches and merging the results produces the largest yield.

Page processing and identification of false positives entries

The majority of disease-based Wikipedia entries follow a predictable page structure; these sections, and their content, can be identified in the plain text returned:

- Signs and symptoms
- Diagnosis
- Risks
- Causes
- Management or Treatment
- Complications

Failure to detect these sections is used indicate a false positive entry. For example, some disorders match one of the 'List of ICD-9 codes' entries (health-related but not suitable here), and some disorders match non-health-related entries ('Burn of elbow' is matched to 'Professional wrestling attacks'). Identifying the entry sections is also important for subsequent processing.

NLP approach

Indexing of SNOMED CT

Using the Natural Language Toolkit [2] active acceptable descriptions of SNOMED CT (including en-US and en-GB variants) were indexed/normalised using the following steps:

- Stop word removal
- ASCII text conversion
- Word tokenisation
- Porter stemming
- Upper case normalisation
- Alphabetic key ordering

Wikipedia entry comparison

Corresponding steps were applied to the text from each Wikipedia entry section, along with sentence tokenisation, word equivalent application and the generation of 1-6 n-gram matching candidates (earlier experiments showed a significant drop in matching for any candidate strings longer than 6 tokens).

Part of speech (POS)-tagging was not used, nor were other sophisticated strategies such as context and negation detection.

The matching approach took advantage of SNOMED CT semantic types, for example only looking for 'Finding', 'Disorder' or 'Event' concepts within 'Complication' sections.

Each n-gram generated was compared with the SNOMED CT index and any matches were stored.

Test data vignette preparation.

For each RefSet member for which a Wikipedia match was identified, test vignettes were prepared using the following approach:

1. As a minimum, each vignette required [1..*] presenting signs or symptoms, [1..1] 'diagnosis' (the name of the disease matched) and [1..*] treatments. This combination supports populating the essential/minimal features of a 'single episode summary', useful for producing coded data for testing discharge summary specifications.

2. Data cleansing and randomisation

Aspects of the NLP pipeline, in particular the use of structural stemming and not using POS-tagging, made the matching process vulnerable to false positives. Manual review of frequently-occurring matches in an initial sample of entries tested allowed genuine false positives to be excluded from subsequent matching. For test data purposes the risks this exclusion approach introduced was felt to be acceptable.

Any direct ancestors or descendants of the disorder concept used for matching were removed.

Optionally each vignette could be based on leaf concepts only (e.g. only using 'abdominal pain' where 'pain' has also been matched) accepting that this reduces the total number of matches available – and thus variety of vignettes.

Depending on the total number of matches in each category, vignettes were then generated using a sample of available matches. If there was only one match then that single value would be used every time.

3. Where available, each vignette could then be augmented with other coded features, categorised according to their chapter locations within SNOMED CT (based on simple inclusion/exclusion constraints) and the entry section in they were detected:

-Specific examination results

-Imaging, Microbiology or Other investigations

-Investigation results

-Associated conditions

-Causal organisms

-Risk factors

-Complications

-Treatments/interventions can be subdivided into:

-Surgical, non-surgical treatments

-Medicinal products (or very rarely substances where the product is not matched but a valid active ingredient can be). Specific 'infective vignettes' can be identified where the causal organism is a bacterium and can be 'matched' to a corresponding antibacterial etc.

As an example, and based on analysis of the Wikipedia page for Mitral valve regurgitation [3], the following sample coded vignette was generated (codes not shown to ease readability):

-Symptoms and signs:

Oedema, Regurgitation, Paroxysmal nocturnal dyspnoea, Heart irregular, Tachycardia, Retching and Orthopnoea

-Specific examination results:

Pansystolic murmur, Third heart sound and Late systolic murmur

-Associated conditions:

Congestive heart failure, Atrial fibrillation and Rupture of chordae tendineae

-Imaging investigations:

Magnetic resonance imaging (MRI) of vessels, Transoesophageal echocardiography, Colour doppler ultrasound and Angiocardiography

-Other investigations:

Electrocardiographic monitoring

-Diagnosis:

Mitral valve regurgitation

-Medication:

Hydralazine, Diuretic and Angiotensin-converting enzyme inhibitor agent

-Non-surgical interventions:

Stretching and Perfusion

-Surgical interventions:

Ring annuloplasty, Construction of shunt and Repair of mitral valve

Quantitative results

Of the 4325 members of the July 2018 GP/FP RefSet:

1588 (37%) match a Wikipedia page satisfying minimum vignette criteria.

Matches were made to 923 unique Wikipedia entries

577 Wikipedia entries matched a unique RefSet member

By contrast, for example 10 RefSet members are matched to the entry for Chlamydia

Qualitative evaluation

In terms of clinical plausibility, synthetic ‘clinical vignettes’ produced this way can reasonably be criticised – some of the coded content does not withstand even casual clinical scrutiny and each ‘story’ implicitly compresses what may be years of care into a short episode. However, even such a ‘fiction’ could still provide SNOMED CT-encoded sample data suitable for several purposes.

Observations and next steps

The vignettes produced so far are clearly far from perfect but demonstrate that even a trivial NLP pipeline can be used to identify SNOMED CT-codable content in textual resources. Performance could clearly be improved by many steps such as use of more sophisticated NLP techniques. What follows are a number of possible uses of data produced this way:

Testing communication standards during their development.

Declared as the original goal of this work, the richer and more varied the payload used in testing a clinical communication specification, the better the testing process. Hand-written test data is time-consuming to produce, and may find itself being reused and therefore systematically fail to detect certain error types.

Putting the 'clinical' back into clinical terminology

SNOMED CT's way of representing and organising its content is valuable – distinguishing content for the above 'vignette' would not have been possible without it, but these organising principles do not 'speak' to all users. SNOMED CT's ideal is stable compartmentalisation of ontologically distinct ideas. The ability to generate 'clinical stories' beyond the immediate neighbourhood of a disease animates SNOMED CT's content, illustrating its relevance to the dynamics of normal clinical activity.

Improving content development processes

The work presented uses and tests features of the SNOMED CT data, which will in turn generate exceptions and errors: some of these errors will be because the process is wrong, but some will be because the data is wrong, and such errors can be reported back to SNOMED CT developers for correction.

By shining a light on relatively neglected parts of the terminology, it is possible that errors in long-standing content will be revealed.

It may also be possible to generate code sets that complement those already being developed. If a clinical group develops a list of diagnoses relevant to their practice, currently that is all they will receive. It may be possible to use a similar process to that described above to produce candidate sets of corresponding treatments, symptoms, causes and complications, thus simplifying, accelerating and enhancing current development processes.

Reference Documentation

[1] https://www.mediawiki.org/wiki/API:Main_page + https://pypi.org/project/wikipedia/

[2] https://www.nltk.org/

[3] https://en.wikipedia.org/wiki/Mitral_insufficiency