Time-sensitive inventory of medical terminology

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

This inventory contains a set of terms that are relevant to the study of medical history.

The inventory is organised as a set of "heading terms" (around 175,000). Each heading term, which belongs to one or more of seven different semantic categories (shown in Table 1), is accompanied a set of semantically-related terms. These related terms have been automatically extracted using text mining methods from large collections of published medical text, dating from 1840 to the present day.

The nature of the semantic relationship holding between the heading term and each related term varies. Some examples of possible relationships holding between pairs of terms include the following:

- **Related terms may be synonyms of each other.** For example, *pulmonary tuberculosis, pulmonary phthisis* and *tuberculous consumption* are identified as terms related to *pulmonary consumption*.
- **One term may be more or less specific than the other**. For example, the terms *flat, shelter, chapel* and *hut* are identified as terms related to *building*.
- **One term may correspond to a part of the other**. For example, *ankle, shin* and *thigh* are identified as terms related to *leg*.
- **One term may occur (spatially) in the proximity of the other**. For example, *larynx, pharynx* and *bronchiole* are identified as terms related to *trachea*.
- **One term may be used in the treatment of the other**. For example, the drugs *glibenclamide, metformin* and *tolbutamide* are identified as terms related to *diabetes*.

The unique feature of our terminological inventory is that the semantically-related terms may correspond to terms used within different periods of time, and which may not be in common usage today.

Over time, shifts/evolutions in terminology, advances in medical knowledge, living conditions, etc., mean that closely related terms are likely to be subject to change over time. Accordingly, in studying historical change, it can often be difficult to know which query terms to use in a search system, to ensure that relevant documents are retrieved from different periods of time.

It can be extremely useful if the system is able suggest terms that are related to query terms, and which help to widen the scope of the search. This may help either in retrieving documents from a wider time range and/or to help them to explore areas related to their original concepts within a particular time period.

As an example, within our inventory, the heading term *rubella* (a viral infection formerly common in children) includes amongst its semantically related terms a

historically-relevant synonym (*Rotheln)*, as well other viruses and viral infections (e.g., *smallpox, chickenpox, rotavirus, poliovirus*), some of which are also particularly common in childhood.

As a further example, the environmental factor *overcrowding* has related terms corresponding to other poor living conditions that may have contributed to certain diseases in the past (*dilapidation, ventilation, cleanliness*, etc.) together with other environmental entities representing the structures in which such conditions occur (e.g., *house* and its more temporally-sensitive synonym *dwelling*, *workroom,* etc.)

1.2 Representation of the lexicon (flat files, database, markup)

The time-sensitive terminological inventory is made available in the format used by the Open Biological Biomedical Ontologies (OBO) (https://oboformat.googlecode.com/svn/trunk/doc/GO.format.obo-1_4.html). This is a widely used format used to encode various different ontologies that cover medical and biomedical subdomains, including diseases, anatomical entities and environmental entities. Releasing the inventory in this format has several advantages. Firstly, using a standardised format opens up possibilities for future integration of the inventory with other ontologies covering relevant areas. Secondly, the files can be processed/visualised with existing tools, such as Protégé (http://protege.stanford.edu/) and OBO-Edit (http://oboedit.org/).

NOTE: Due to the large size of some of the terminology files, the default memory allocation for these tools must be changed. We found that Protégé should be used preferentially for visualising the files, since it seems better able to handle the large files, especially when the memory allocation flags set in the *run.command* file are set to *-Xmx6G -Xms2G*. Whilst OBO-Edit also seems able to load the larger files (especially when the OBOEdit.vmoptions file is edited to allocate more memory, we tried -*Xmx6000*M), it seems less responsive than Protégé when a large file has been loaded.

1.3 Character encoding

The characters have been encoded in UTF8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sophia Ananiadou
Address: Manchester Institute of Biotechnoogy,131 Princess Street, Manchester M1 7DN, UK
Affiliation: National Centre for Text Mining, School of Computer Science, University of Manchester
Position: Professor
Telephone: +44 161 306 3092
Fax: +44 161 306 5201
e-mail: Sophia.Ananiadou@manchester.ac.uk

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available as an archive from the META-SHARE platform.

2.3 Copyright statement and information on IPR

The resource is licensed under a Creative Commons Attribution licence (CC-BY). If you use the resource, please attribute the National Centre for Text Mining (NaCTeM), School of Computer Science, University of Manchester.

3. TECHNICAL INFORMATION

3.1 Directories and files

The archive contains the directory *medical_inventory*. The terms in the inventory are split into separate files according to the semantic category/ies to which the heading term belongs (see Table 1 for detailed information). The following files are in the directory:

- *all_nes_bmj_moh_normalised.thesaurus_anatomical_entity.obo* – File containing heading terms categorised as anatomical entities.
- *all_nes_bmj_moh_normalised.thesaurus_biological_entity.obo* - File containing heading terms categorised as biological entities,
- *all_nes_bmj_moh_normalised.thesaurus_condition.obo* - File containing heading terms categorised as medical conditions.
- *all_nes_bmj_moh_normalised.thesaurus_enviromental_entity.obo* - File containing heading terms categorised as environmental entities.
- *all_nes_bmj_moh_normalised.thesaurus_sign_or_symptom.obo* - File containing heading terms categorised as signs or symptoms
- *all_nes_bmj_moh_normalised.thesaurus_subject.obo* - File containing heading terms categorised as subjects
- *all_nes_bmj_moh_normalised.thesaurus_therapeutic_or_investigation_entity.obo* - File containing heading terms categorised as therapeutic or investigational entities.
- *medical_inventory_README.txt* – Contains a description of the resource, an overview of the methods used to create it, a description of the format of the entries, etc.
- *medical_intentory_licence.txt* – Provides information about the licence assigned to the medical inventory.

3.2 Data structure of an entry

The inventory is organised as a set of "heading terms". Each heading term corresponds to a named entity (NE), belonging to one of the seven semantic categories shown in Table 1.

The heading terms used are those obtained an named entity (NE) recognition tool over two large archives of published medical text (see below), spanning a long period of time (1840 – 2013). The complete archives were then further processed to determine the various textual contexts in which each of the heading terms can appear; the 20 terms which appear in the most similar contexts to the heading term were extracted as related terms and accompany each heading term.

Table 1: Types of terms included in the medical inventory

Entity Type	Description	Examples
Condition	Medical condition/ailment	phthisis, bronchitis, typhus
Sign_or_ Symptom	Altered physical appearance/behaviour as probable result of injury/condition	cough, pain, rise in temperature, swollen
Anatomical	Entity forming part of human body, including substances and abnormal alterations to bodily structures	lung, lobe, sputum, fibroid
Subject	Individual or group under discussion	children, asthma patients, those with negative reactions to tuberculin
Therapeutic_or _Investigational	Treatment/intervention administered to combat condition (including diet/foodstuffs), or substance, medium or procedure used in investigational medical or public health context	atrophine sulphate, generous diet, change of air, lobectomy
Biological Entity	Living entity not part of human body, including microorganisms, animals and insects	tubercle bacilli, mould, guinea-pig, flea
Environmental	Environmental factor relevant to incidence/prevention/control/treatment of condition. Includes climatic conditions, foodstuffs, infrastructure, household items or occupations whose environmental factors are mentioned	humidity, high mountain climates, infected milk, linen, drains, sewers, dusty occupations

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The time-sensitive terminological inventory contains about 175,000 heading terms. The approximate size of the files containing the terminological data is a containing the inventory (*all_nes_bmj_moh_normalised.thesaurus.obo*) is as follows:

- *all_nes_bmj_moh_normalised.thesaurus_anatomical_entity.obo* – 58MB
- *all_nes_bmj_moh_normalised.thesaurus_biological_entity.obo* - 7MB
- *all_nes_bmj_moh_normalised.thesaurus_condition.obo* – 71MB
- *all_nes_bmj_moh_normalised.thesaurus_enviromental_entity.obo* – 54MB
- *all_nes_bmj_moh_normalised.thesaurus_sign_or_symptom.obo* – 39 MB
- *all_nes_bmj_moh_normalised.thesaurus_subject.obo* - 11MB
- *all_nes_bmj_moh_normalised.thesaurus_therapeutic_or_investigation_entity.obo* – 78MB

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the medical terminological inventory is English.

4. 2 Entry Type

In the final inventory, each heading term is accompanied by the following information:

- The set of the 20 terms that are considered to be most closely related to the heading term, according to contextual similarity.
- Each of the 20 most related terms is accompanied by a numerical score that represents the degree of similarity between the contexts of the related term and the contexts of the heading term. This score is called the cosine similarity.
- The semantic category (amongst those listed in Table 1) which was assigned to the heading term by the NE recognition tool
- Links to one or more concepts in the UMLS Metathesaurus ((https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html). This is another large-scale resource that includes biomedical and health related concepts, which has been largely created manually. In the Metathesaurus, concepts are represented by a set of synonymous terms, and different types of relationships are identified between concepts. In our terminological inventory, we include the identifiers of all Metathesaurus concepts in which our heading term figures within their list of synonyms. Whilst there is a large degree of overlap between the scopes of the UMLS Metathesaurus and our terminological inventory, it is expected that the information present within each resource can complement the information present within the other. Specifically, the UMLS Metathesaurus does not aim to include comprehensive coverage of historical term variants, whereas the identification of historical related terms has been a major aim of our work. Thus, the inclusion of potentially related UMLS concept identifiers within our inventory provides future scope for linking together the two resources.

.

4.3 Attributes and their values

An example entry in the OBO files (for the term pulmonary tuberculosis) is shown below.

```
[Term]
id: HOM:1159
name: pulmonary tuberculosis
synonym: "non pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.9285235631916715]
synonym: "miliary pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8890370477480192]
synonym: "tuberculosis pulmonary" RELATED DS_RELATED [DS_SCORE:0.8694139206266057]
synonym: "chronic pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8654668011030042]
synonym: "pulmonary y tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8619417498619306]
synonym: "pulmonary phthisis" RELATED DS_RELATED [DS_SCORE:0.8421124650540113]
synonym: "inactive pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.832065670929346]
synonym: "old pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8307341025365899]
synonym: "primary pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8298542706890558]
synonym: "bilateral pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8191471054604339]
synonym: "tuberculosis miliary" RELATED DS_RELATED [DS_SCORE:0.8123758775790052]
synonym: "incipient pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8069857568206482]
synonym: "pulmonary mycobacterium tuberculosis" RELATED DS_RELATED
[DS_SCORE:0.8068339428878909]
synonym: "pulmonary tuberculosis active" RELATED DS_RELATED [DS_SCORE:0.8064070676210551]
synonym: "infectious pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.80465669670636]
synonym: "extrapulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.8028301888191761]
synonym: "acute miliary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.7976048542827227]
synonym: "symptomless pulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.7963059836059512]
synonym: "chronic miliary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.795678270460759]
synonym: "nonpulmonary tuberculosis" RELATED DS_RELATED [DS_SCORE:0.7922752141018254]
subset: condition
xref: UMLS:C0041327
```

The general format of OBO files is fully documented elsewhere, but here we explain some of the features of the OBO file encoding the time-sensitive medical terminological inventory.

Each entry includes the following parts:

- **[Term]** - The first line of an entry.
- **id** - provides a unique id for the term
- **name** - the "heading term" for this entry
- **synonym** - denotes one of the 20 more similar semantically related terms for each heading term, according to distributional semantics. Note that the semantically related term is not necessarily a synonym, but the word is required to comply with the OBO format; a more exact characterisation of the nature of the relationship is provided in a further part of the line, as detailed below. Following the hyphen, there are four further parts to each line, separated by spaces:
    - **The related term**, enclosed in quotes
    - **RELATED** - One of a fixed set of synonym scope values possible in the OBO format, denoting that there is a semantic relation of a general nature between the heading term and the semantically-related term. This is used in all cases, since current process for extracting terms does not attempt to distinguish between different types of semantic relations.
    - **DS_RELATED** - denotes that the related term has been extracted using distributional semantics techniques.
    - **DS_SCORE** - A score for the related term. This score (the cosine similarity) represents the level of similarity between the textual contexts of the head term and the textual contexts of the related term.
- **subset** - one or more lines corresponding to the NE categories that were assigned to the term in the BMJ and MOH archives after running the NE recogniser over them. Possible values are as follows:
    - *anatomical_entity*
    - *biological_entity*
    - *condition*
    - *enviromental_entity*
    - *subject*
    - *sign_or_symptom*
    - *therapeutic_or_investigation_entity*
- **xref** - zero or more lines with an identifier (beginning with a "C"), corresponding to a concept identifier in the UMLS Metathesaurus (https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html).

4.4 Coverage of the lexicon

The inventory was created automatically by applying a text mining technique called distributional semantics to two large collections of published medical documents, each spanning a long period of time. These collections are as follows:

- The archive of the British Medical Journal (BMJ) (http://www.bmj.com/archive). This journal is aimed at medical professionals, and includes various types of

articles, including research, analysis, practice, case reports, letters, and obituaries. We worked with a collection of approximately 380,000 articles, spanning from 1840 to 2013.

- The London Medical Officer of Health reports (MOH) reports, digitised by the Wellcome Library (http://wellcomelibrary.org/moh/) are concerned with examining public health issues in different London boroughs. The archive consists of around 5,000 reports produced between 1848 and 1972, whose lengths range from a few pages to several hundred pages.

Distributional semantic models (DSMs) exploit the observation that words that appear in similar contexts often exhibit similar meanings. The archives were processed to determine the various textual contexts in which each of the heading terms can appear; By finding terms with similar contexts over the complete archives, we try to ensure that related terms that may be relevant at different periods of time are included in the resource.

4.5 Intended application of the lexicon

The time-sensitive inventory of medical terminology is primarily intended to be used in search systems over historical medical archives, as means to help users to widen the scope of their search, both in terms of the range/depth of topics explored, and in order to allow the retrieval of relevant documents over a wide time period.

When a term is searched for by a user, it can be looked up as a heading term in the terminological inventory, and the 20 most related terms listed can be suggested as possible ways to expand the query, as is done in the History of Medicine semantic search system (http://www.nactem.ac.uk/hom/). The inclusion of the contextual similarity scores for each related term also allows the possibility of filtering the complete list of related terms according to some threshold, such that less related terms are excluded. Furthermore, linking with the UMLS Metathesaurus could help to identify further potentially related terms.

The screenshots below show how the terminological inventory is used in the History of Medicine semantic search system.

When a medically-relevant search term is entered by the user, a set of related terms is displayed, by accessing the time-sensitive terminological inventory. This is illustrated in Figure 1. The differing sizes of the related terms provide an indication of their level of contextual similarity to the query term entered, found by accessing the scores associated with the related terms. In the interface, clicking on a term causes it to be added to the query.

Figure 2 shows how the frequencies of occurrence of the terms in the document archives over time are used as a guide to indicate the time-senstive nature of the terms. In this example, the user has widened their search by choosing the term *pulmonary phthisis* as a related term of *pulmonary tuberculosis*. The graph shows that *pulmonary phthisis*, whilst initially more common than *pulmonary tuberculosis*, largely fell out of use after the 1930s.

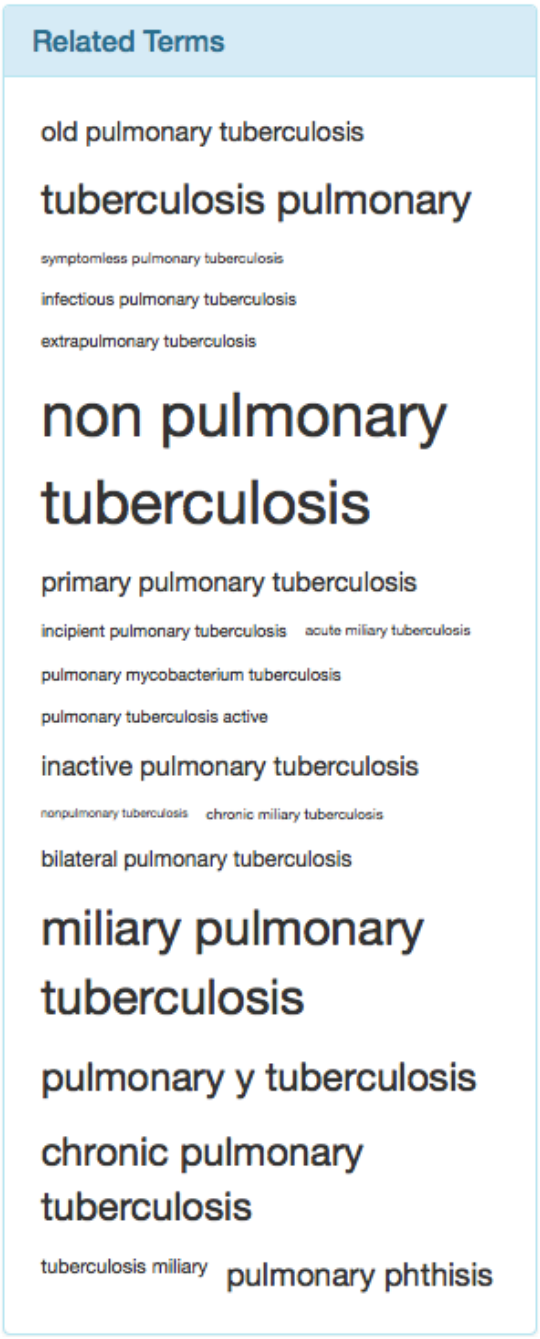

Figure 1: Display of related terms in the HOM interface

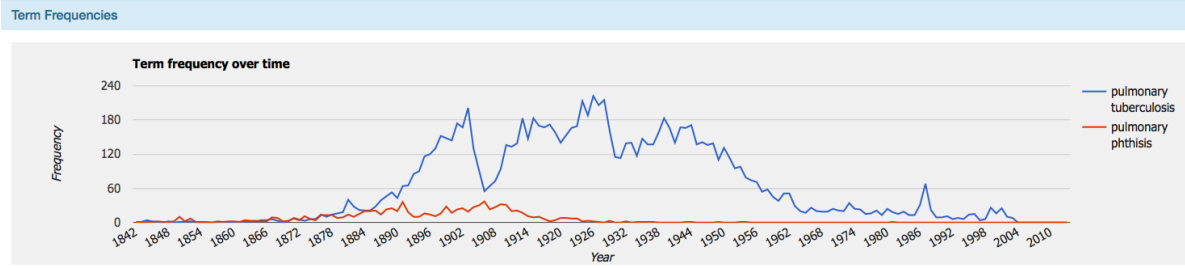

Figure 2: Term usage over time in the HOM interface.

4.6 Reliability (automatically/manually constructed)

The medical inventory is constructed using fully automatic methods. An evaluation of the performance of our method in terms of its ability to recognise related terms of disease entities showed that it was able recognise synonyms that are not listed in the UMLS Metathesaurus and that the majority (62%) of automatically identified terms was deemed to be semantically related to the relevant heading terms.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Thompson, P., Carter, J., McNaught, J. and Ananiadou, S. (2015). **Semantically Enhanced Search System for Historical Medical Archives**. In *Proceedings of DigitalHeritage 2015*

The inventory is introduced in the following article:

Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys and Sophia Ananiadou (2015). **Text Mining the History of Medicine**. *PLOS ONE*.