KConnect UMLS Annotation Task Manual

June 2016

Introduction

As patient records are increasingly in electronic form, efforts are being made to utilize these data to support medical research, for example into drug adverse reactions, and for clinician support, for example presenting the patient record in more informative ways. An important building block for this work is the ability to automatically (using computerized methods) recognise the appearance of concepts in free text notes. Concepts of interest might be the names of drugs, diseases, symptoms or anatomical parts. Identifying the appearance of concepts is complicated by the fact that the same word can mean different things. For example, the word “hand” might refer to the body part, or to the handing over of something, among others. Our research focuses on being able to do this disambiguation task well, and the methods we employ require data in the form of examples that have been disambiguated by humans with the right expertise, in order to use learning methods to enable the computer to make the same choices. This is why we are asking you to perform this task of choosing the right interpretation for concepts mentioned in clinical text.

You will be presented with a series of sentences from patient records. In the sentence, a mention will be highlighted. Underneath, you will see a series of alternatives for what that mention might refer to. Your task is to pick the one that you think is the best fit (or reject them all). You will then advance to the next mention, which might be later in the same sentence or might be in the next sentence in the record, or the next record.

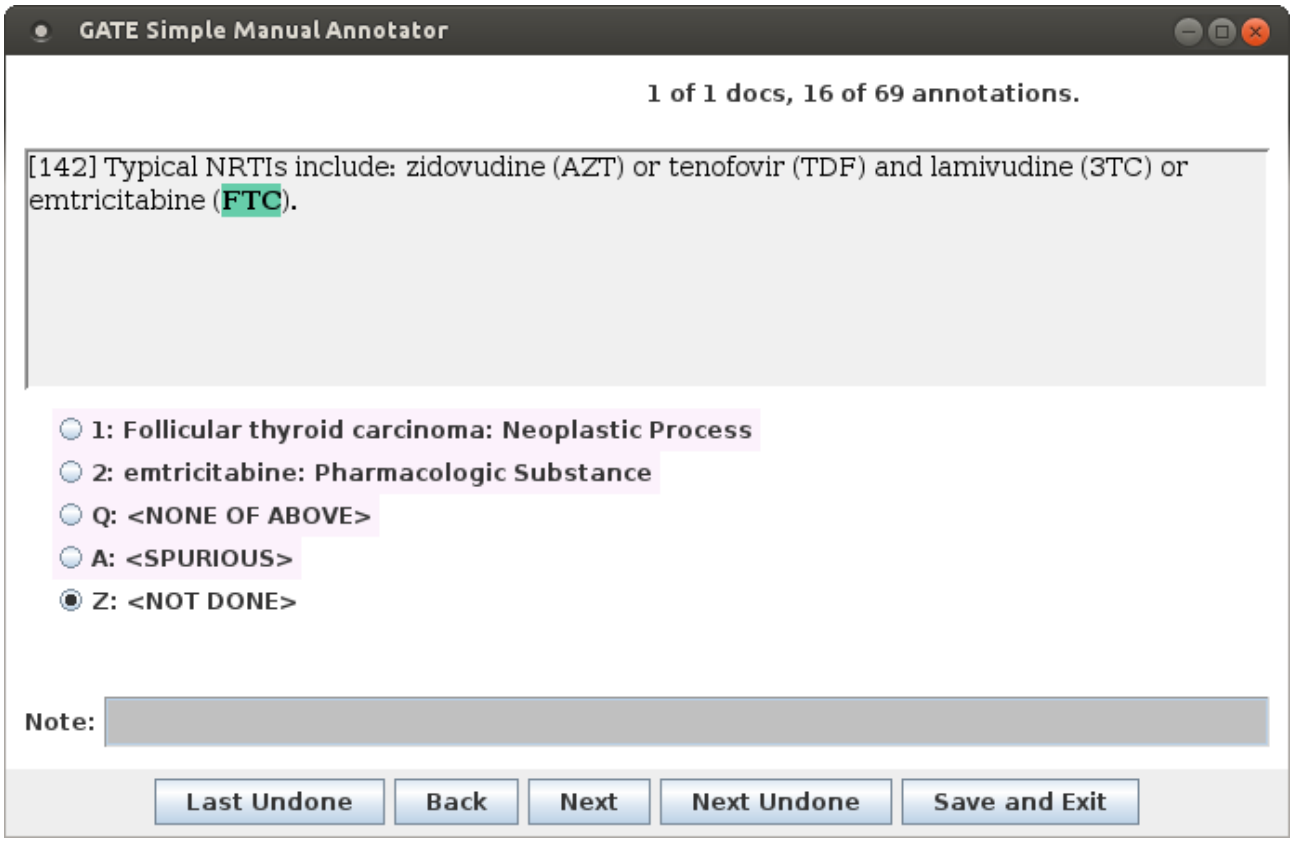
The text you will see is from patient records that have been anonymized. Nonetheless the text is highly confidential, and you are reminded that you have agreed to conditions regarding what constitutes a valid justification for your viewing of the records; in our case to conduct research only. You will access the records via a remote desktop to an NHS computer. The patient data must remain on the NHS computer, and under no circumstances be transferred in any part to another computer. **You must not copy and paste text out of this remote desktop to the local computer, or in any other way take the data off the NHS computer. If this happens for any reason, you must notify us immediately.**

This annotation project is being run as part of our work on the KConnect project (grant agreement No 644753, EU Horizon 2020). The data you will view are a selection from the CRIS database hosted at South London and Maudsley NHS trust (http://www.maudsleybrc.nihr.ac.uk/about-us/core-facilities/clinical-record-interactive-search-cris/). Your contacts with regards to this work are Dr Angus Roberts, Dr Genevieve Gorrell and Mr Xingyi Song.

The next section describes the use of the annotation tool that you will use to perform the task. The following section describes how to correctly select the right interpretation in various cases.

Using the annotation tool

The task facilitator will set up the tool ready for you to use. The tool will automatically find the next un-done mention in the corpus you are assigned. The position you are currently at in the corpus is indicated at the top. Firstly, the current document and the number of documents in your corpus are indicated. In the screenshot, we are on the first of one document (“1 of 1 docs”). Secondly, the current mention and the number of mentions in the current document are indicated. In the screenshot, we are on the 16th mention of 69 (“16 of 69 annotations”). The mention is highlighted in teal and the context is presented surrounding it.



The choice list is presented below the mention. Numbers are assigned to each choice, and letters are assigned to the options. This means you can use key presses to annotate quickly, by pressing the number or the letter for that option. If you want to use the number pad on the right of the keyboard, you need to press "Num Lock" on your keyboard first. You are of course welcome to use the mouse to select your option too. If keyboard presses don't seem to be working, you might need to first click in the window to give focus to the tool. If you are having problems, ask your task facilitator.

Navigation options allow you to move forward and back through the mentions, or jump to the previous (last) or next un-done mention. You can also use right or down arrows to move to the next mention, or left or up arrows to move to the previous one. If you are using the autoadvance option, when you select a choice the tool will automatically move to the next mention, so you may never need to use the navigation buttons. The autoadvance option is faster but can be disconcerting. You may discuss your preference with the task facilitator, who can change the setting.

A "Save and Exit" button is provided for your reassurance, but in fact closing the window will also save your work. Additionally every time you move to a new document (as indicated at the top of the

window, by having moved to “1 of ... annotations”), your work on the previous document is saved. If your computer crashed, you would lose the work on the current document, but not on completed documents. There is no way to not save your work on exit. However you can quickly remove a large number of annotations by holding down Z with autoadvance switched on.

We can only make use of completed documents, so please try to finish the document you are working on before finishing, i.e. stop when you have a “1 of ... annotations” at the top.

Performing the disambiguation

The task is to identify which of several candidates is the best match to an occurrence in text. The candidates have been prepared automatically from a variety of vocabularies which combine to create a list. As a result of this automatic method, sometimes the options are a bit odd or underspecified, and sometimes the correct option is not available. For the most part, however, it should be possible to select the correct option. Sometimes only one option is available, and that option is obviously correct. We still need you to confirm that it is correct by selecting it.

Each option concludes with a colon and then a category, as in, for example “Acquired Immune Deficiency Syndrome: Disease or Syndrome”, showing that the concept indicated, referred to by the name “Acquired Immune Deficiency Syndrome” is a disease or syndrome. This is intended to make it easier for you to understand what that option is, which isn't always otherwise clear. In the case of the mention “fall” in the following sentence, for example, there might be this option “1: Falling: Finding”.

“Once levels **fall** below 50 copies/mL checks every three to six months are typically adequate.“

The inclusion of “Finding” enables us to interpret it. It would otherwise be unclear what “Falling” means in that option. A finding is a medically relevant observation about a patient, and so “Falls” probably refers to the finding that the patient falls over, and in this case isn't the correct interpretation. In this case, “fall” is a quantitative concept. Even if “fall” appears in the context of a medically relevant observation, such as “ Mrs ZZZZZ’s potassium levels are falling dramatically”, the mention, spanning only the word “falling” would still be a quantitative concept, since only the longer span of “potassium levels are falling” would be a medically relevant observation. “Falling” itself simply describes the relevant quantitative concept relating Mrs ZZZZZ's potassium levels to their previous condition.

As a rule, you should select the entity that matches what the highlighted span of text refers to only. For example, the phrase “acute anterior wall myocardial infarction” might appear in text. If only “myocardial infarction” is highlighted, you shouldn't select the candidate for acute anterior wall myocardial infarction, even though you know that in this case, that is what it was.

Mentions we aren't interested in: when to select “spurious”

We are only interested in the following categories of mention:

- ✓ **Anatomy**: including all aspects of the anatomy and also all abnormalities.
- ✓ **Diseases and symptoms**: including injuries, accidents and psychological disorders.
- ✓ **Investigations**: laboratory tests etc.
- ✓ **Drugs:** legal and illegal, and pharmacologically active substances.
- ✓ **Observations:** all findings potentially relevant to a disorder, for example falling, or having a family history of a condition, or being pregnant.
- ✓ **Health care organizations:** such as a care home, hospital or the NHS
- ✓ **Therapies, treatments and healthcare activities:** such as taking blood
- ✓ **Time expressions:** last week, never, usually, for two weeks etc.

In the case that a mention does not fall into one of these categories, you should select “spurious”. For example, “falls” in the above example referred to a decline in quantity, and is spurious, because it doesn't fall into the four categories listed above. In the context of “Mrs ZZZZZ experienced a number of falls last year”, this is medically relevant and “falls” in this case is **not** spurious.

The mention “saw” in the case of “I saw her for the first time last week” isn't medically relevant, and is spurious, but in the case of “Mrs ZZZZZ saw for the first time after successful eye surgery,” it is a medically relevant finding.

The categories of interest are more fully explained in the glossary below. You should familiarize yourself with the list so that you are able to recognise when a mention is or is not spurious.

The right answer isn't available, but the mention is of a category of interest: when to select “none of the above”

If the right answer isn't on the list, but the mention refers to an entity of a type that we are interested in, then you can select “none of the above” and also make a note in the notes field at the bottom of what the correct interpretation should be. This will enable us to look it up later.

What to do if two interpretations are equally correct

Where a more general interpretation answers the case, that is better than a more specific one. For example, don't select “peripheral blood” where “blood” is equally fitting. Where one option is in plain English but the other is a more technical term, if both are accurate, choose the one in plain English. For example, don't choose “quetelet index” if “body mass index” is available as an option.

In the case that two correct options are synonyms and both in your judgement equally acceptable, please pick the shortest. In the following example, we have alternatives for "birth defects":

1: Congenital Abnormality: Congenital Abnormality
2: Deformity: Congenital Abnormality

In context: "Certain medications may be associated with **birth defects** and therefore may be unsuitable for women hoping to have children."

In this case, if you consider that 1 or 2 are equally acceptable alternatives, you pick the second because it is shorter.

What to do if it is unclear what an option refers to

We have tried our best to ensure that the option presented to you is of good enough quality to enable you to know what it means, but may not always have succeeded. In the case of, for example, “related”, in the following sentence, the option presented is “1: Related: Finding”, which isn't very clear.

“Specific adverse events are **related** to the antiretroviral agent taken.”

Usually where the option is unadorned by further specification, it is reasonable to assume that it refers to the most obvious, general meaning of the term in a medical context. In this case, we also know that it is a “finding”, and therefore refers to a medically relevant observation about a patient. This suggests that “1: Related: Finding” refers to the patient's being related to other persons, which in this case isn't the right answer.

If you can make a reasonable deduction in such cases, please go ahead and do so, as our aim is to get as much data as possible of a good enough quality rather than to aim for perfection. However if you really don't know, select “none of the above” and write in the notes field what it should be.

Glossary

Annotation (noun): a section (span) of text to which information has been added, for example to indicate that the span of text is the name of a drug.

Annotation (verb): the activity of marking spans of text with information such as the type of entity the text refers to.

Candidates: the different possibilities for what a mention might be referring to.

Corpus: the set of documents (patient records) that we are working with with.

Entity: the real world object that a mention in text might refer to. For example, the word "paracetamol", appearing in text, most likely refers to the real world pharmacologic substance also known as acetaminophen, though could conceivably not do.

GATE: a Sheffield University computer science department software project providing the tools for natural language processing that we are using in the KConnect project.

Local computer: the computer you are currently working on.

Mention: the section of text that refers to an entity.

Remote desktop: the presentation on your computer of the desktop environment of another computer, allowing visual display, mouse interaction etc. as though you were working directly on that computer.

Spurious: an annotation that should not be there, because it is not the kind of mention we are interested in.

Term: a name for something.

Vocabularies (medical): an electronic resource listing terms (names) and the entities to which they may refer. For example, SNOMED CT is a vocabulary.

Types

Anatomy

- **Body Location or Region**—body areas such as shoulder, or neck of the bladder
- **Body Part, Organ, or Organ Component**—body parts or organs, such as the bladder
- **Body Space or Junction**—body spaces and interstices such as the temporomandibular joint
- **Body System**—functionally defined anatomical concept, such as the vascular system
- **Tissue**—body tissues, for example, smooth muscle tissue
- **Anatomical Structure**
    - **Anatomical Abnormality**—abnormalities such as lordosis or septal defects
        - **Acquired Abnormality**—acquired anatomical abnormalities such as intestinal perforation
        - **Congenital Abnormality**—congenital anatomical abnormalities such as fetal methotrexate syndrome, or dyschondroplasias.

Disease

- **Injury or Poisoning**—for example, cerebral trauma or food poisoning

- **Pathologic Function**—for example, rhabdomyolysis or hyperplasia
  - **Disease or Syndrome**—such as infectious diseases and infections
    - **Mental or Behavioral Dysfunction**—for example substance use disorders, aphasia
    - **Neoplastic Process**—cancer
  - **Experimental Model of Disease**—these are models used to investigate disease processes, such as experimentally induced tumours and gene knockout models
  - **Cell or Molecular Dysfunction**—e.g. chromosome deletion or trisomy

Investigation

- **Diagnostic Procedure**—biopsies, mammography, etc.
- **Laboratory Procedure**—haematology, cultures etc.
- **Laboratory or Test Result**—for example, the finding of E. coli, or ECG finding of myocardial infarction
- **Research Activity**—this covers research-related concepts and activities, such as randomization or the use of the qualitative method
  - **Molecular Biology Research Technique**—for example mutagenesis or cloning

Drug

- **Clinical Drug**—vaccines, multivitamins. Clinical drugs are productized/manufactured drugs.
- **Pharmacologic Substance**—substances with pharmacological relevance, such as plant extracts, elements etc.
  - **Antibiotic**—amoxycillin etc.

Observation

- **Finding**—this is an extremely broad category describing all medically relevant observations. Examples might include a fetus being small for gestational age, a patient's history of factory work or a family history of drug abuse. Often, the same concept can be a finding in one circumstance and not in another. For example, a patient might be described as angry in the context of their appointment being cancelled, with no implication that this is medically relevant, or they might be described as angry as part of an assessment at a psychiatric consultation. In the former case, this wouldn't be a finding, but in the latter it would.
  - **Sign or Symptom**
- **Individual Behavior—**for example, smoking,
- **Clinical Attribute**—similar to finding.

Health Care

- **Health Care Activity**—an activity of or relating to the practice of medicine or involving the care of patients. For example, taking blood, or sectioning.
- **Health Care Related Organization**—an established organization which carries out specific functions related to health care delivery or research in the life sciences.
- **Self-help or Relief Organization**—an organization whose purpose and function is to provide assistance to the needy or to offer support to those sharing similar problems.
- **Therapeutic or Preventive Procedure**—a procedure, method, or technique designed to prevent a disease or a disorder, or to improve physical function, or used in the process of treating a disease or injury.

Temporal Concept*—a* concept which pertains to time or duration.