SNOMED Mapping Guidelines for Thai Healthcare NER Dataset

This document outlines the methodology used to annotate Thai healthcare text with SNOMED CT concept types as part of the Thai Healthcare NER dataset construction. The annotation process involved the following steps:

1. **Concept Selection**
    - We used the 19 top-level SNOMED CT hierarchies as the initial concept set.
    - Only 18 types were used in practice, excluding " SNOMED CT Model Component " due to irrelevance in text-based NER.
    - The list of the selected Top Level Concepts with a brief description of the content is represented in below able

Top Level Concepts	Description
Body structure	Represents normal and abnormal anatomical structures such as the mitral valve structure and adenosarcoma.
Clinical finding	Represents clinical observations, assessments, or diagnoses such as asthma, headache, and normal breath sounds.
Environments and geographical locations	Includes physical environments and named places such as countries, regions, or care settings (e.g., intensive care unit, Denmark).
Event	Represents occurrences unrelated to procedures, such as floods and earthquakes.
Observable entity	Refers to things that can be measured or observed, for example, systolic blood pressure and gender.
Organism	Includes organisms relevant to healthcare, such as Streptococcus pyogenes and the domain Bacteria.
Pharmaceutical/biologic product	Refers to drug products and related items, for example, amoxicillin 250mg capsule).
Physical force	Represents forces that may cause injury, for example, friction and radiation.
Physical object	Includes man-made or natural physical items, for example, a vena cava filter or an automobile.
Procedure	Represents medical activities, including surgeries, imaging, or therapies such as appendectomy and physiotherapy.
Qualifier value	Values used to modify or qualify other concepts such as "left" and "severe."
Record artifact	Documents or records used in healthcare, for example, patient records and the family history section.
Situation with explicit context	Captures clinical concepts with embedded context, such as history of myocardial infarction.
SNOMED CT Model Component*	Contains technical metadata supporting the SNOMED CT release.
Social context	Social or cultural factors relevant to care, such as occupation and religious belief.
Special concept	Includes non-logical or navigation-based terms, for example, alternative medicine poisoning.
Specimen	Items collected for testing or analysis, such as urine specimen.
Staging and scales	Assessment tools and classification systems, for example, the Glasgow Coma Scale and FIGO staging.
Substance	General or chemical substances, for example, insulin, and methane.

2. **Prompt Design for Typhoon 2.0**
    - Each Thai sentence or compound medical term was formatted with an instruction prompt like: **"Annotate the given text in BIO tagging format follow the Top Level Concepts of SNOMED CT. Cover the token with tag of concept: [Thai text]"**

- The expected model output should follow the BIO tagging format, e.g.: **["B-Procedure", "I-Procedure", "O", "B-Substance", ...]**

3. **Model Annotation Process**
    - **Typhoon 2.0** was used to generate concept annotations in response to the above prompts.
    - All model responses were filtered to ensure tags matched only valid SNOMED CT Top Level categories.

4. **BIO Conversion**
    - Span-level annotations returned by Typhoon were converted to token-level BIO tags.
    - Thai tokenization was handled using **PyThaiNLP** to preserve alignment.

5. **Handling Ambiguities and Invalid Tags**
    - If an entity type produced by the model was not in the valid SNOMED CT list, it was defaulted to "O" (non-entity) or flagged for human review.
    - A sample of 50 annotated entries was reviewed by healthcare domain specialists to ensure correctness.

6. **SNOMED CT Hierarchy Mapping**
    - Each predicted tag was mapped to its corresponding SNOMED CT top-level concept.
    - A manually curated lookup table was used to validate each mapping.

7. **Annotation Confidence**
    - Confidence was estimated based on Typhoon's internal response consistency and agreement with human annotation.
    - Manual corrections were applied where model outputs were unclear or ambiguous.

Summary

These guidelines were followed to produce a high-quality, weakly supervised dataset for Thai healthcare NER aligned with international SNOMED CT standards. The methodology supports reproducibility, multilingual adaptation, and downstream tasks such as ontology-grounded language modelling, information extraction, and clinical knowledge graph construction in low-resource settings.