Abs-168

Saman Hina (University of Leeds), Eric Atwell (University of Leeds), and Owen Johnson (University of Leeds)

Enriching a Healthcare Corpus with SNOMED-CT standard medical semantic tags

Clinical patient records include a mix of structured data and narrative text. We have compiled a corpus containing diverse medical narratives from different healthcare partners, and are enriching it with concept-tags from the international standard SNOMED-CT Systematized Nomenclature of Medicine - Clinical Terms ontology. SNOMED CT provides a concept-based classification of clinical terms (Elkin et al. 2006). This paper describes the different partially-structured data sets including doctor's progress notes, hospital discharge summaries, verbal autopsies, and synthetic data from undergraduate and postgraduate medical training. The semi-structured data contains noise and inconsistencies at several levels, such as variable spelling, punctuation and abbreviations of clinical terms, through to wide variations in document structure.

We have developed a prototype medical semantic tagger to annotate the text with semantic tags from the healthcare data standard SNOMED-CT, advocated by the US College of American Pathologists and UK National Health Service to store, retrieve, and standardize clinical information (IHTSDO 2010). The SNOMED-CT ontology includes over 300,000 medical concepts with some very fine-grained distinctions; for example, some of the complex multiword concepts present in SNOMED CT are;

Structure of intervertebral foramen of fifth thoracic vertebra.

Entire pterygoid process of sphenoid bone.

Diagnostic radiography of sacrococcygeal joint.

Structure of venous plexus of the hypoglossal canal.

Structure of posterior temporal diploic vein.

Entire intervertebral disc space of seventh cervical vertebra.

Aliphatic carboxylic acid, C10-C26

Black locks, oculocutaneous albinism, AND deafness of the sensorineural type

The challenge with this ongoing research is to use this data standard in order to extract broader semantic types from the noisy corpus of medical narratives. Previous researchers have investigated biomedical text corpora by parsing it with English language parsers and have improved the accuracy of automatic processing of biomedical texts by simplification of the sentences (Jonnalagadda et al. 2009). Our systems aim to evaluate the complexity of the corpus as well as the SNOMED CT data standard. We have devised a broader medical semantic tag-set, including medical semantic classes such as finding, substance, disorders, sports injury, situation, event etc derived from the SNOMED-CT ontology hierarchies.

For corpus quality assurance, we use GATE - General Architecture for Text Engineering (Cunningham et al. 1997). Our medical tagger will help language and biomedical researchers to annotate clinical texts with authenticated semantic types without relying on annotation guidelines drawn up by domain experts. Our tagger will also help in secure information extraction from clinical documents (Hina et al. 2010) .

References:

Cunningham, H., K. Humphreys, et al. (1997). "GATE - a TIPSTER-based General Architecture for Text Engineering. In the TIPSTER Text Program (Phase III) 6 Month Workshop."

Elkin, P. L., S. H. Brown, et al. (2006). "Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists." Mayo Clinic Proceedings 81(6): 741-748.

Hina, S., Atwell. E., et al. (2010). "Secure Information Extraction from Clinical Documents Using SNOMED CT Gazetteer and Natural Language Processing" Proceedings of The 5th International Conference for Internet Technology and Secured Transactions ICITST-2010

Jonnalagadda, S., L. Tari, et al. (2009). Towards effective sentence simplification for automatic processing of biomedical text. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Boulder, Colorado, Association for Computational Linguistics.