ASE: A Search Engine for Semantically Annotated Documents

Michael Lawley | Group Leader, AEHRC
michael.lawley@csiro.au

Alejandro Metke-Jimenez | Post Doc, AEHRC
alejandro.metke@csiro.au

THE AUSTRALIAN E-HEALTH RESEARCH CENTRE
www.csiro.au

Agenda

- **Introduction**
- Keyword search
- Semantic search
- The ASE hybrid model
- Conclusion

Introduction

- **Searching free-text in the clinical informatics domain is an important and challenging task**
  - In primary health care settings effective search over patient health records may improve health outcomes and save costs
  - Effective search is also useful in tasks such as cohort identification in clinical trials

- **Traditional keyword-based search and semantic search both have limitations**
  - Keyword-based search is affected by synonimity and ambiguity
  - Semantic search relies on semantic annotations which, in most cases, need to be automatically generated

Agenda

- Introduction
- **Keyword search**
- Semantic search
- The ASE hybrid model
- Conclusion

Keyword search

- **Different IR approaches**
  - Standard boolean model
  - Vector space model
  - Language models
- **Several open source search engines available**
  - Lucene
  - SOLR
  - Indri
  - Terrier
  - MG4J

Limitations of keyword search

Document

The patient presented to the ED with ACS and was given Plavix.

Query

Patients with Acute Coronary Syndrome that were given Plavix

- **The acronym “ACS” is a synonym of “Acute Coronary Syndrome”**
  - In the example, these terms would not match the document in a keyword search
- **The acronym “ACS” is also ambiguous**
  - In the medical domain it can also represent “altered conscious state”, “American Cancer Society”, and “American College of Surgeons”

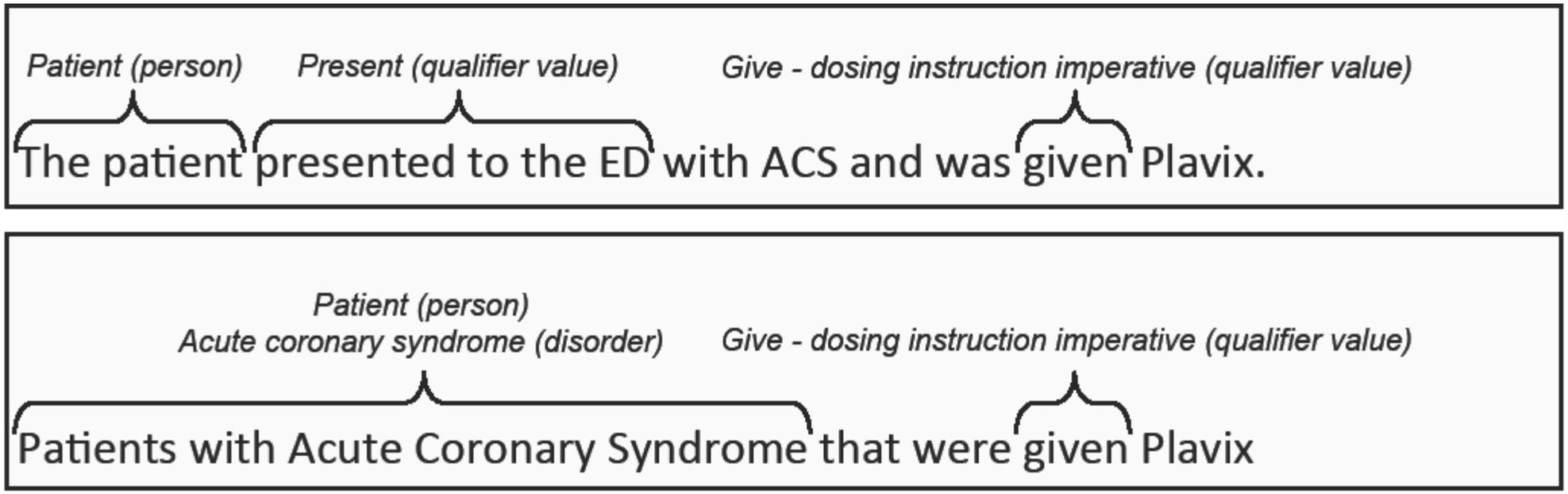
Agenda

- Introduction
- Keyword search
- **Semantic search**
- The ASE hybrid model
- Conclusion

Semantic Search

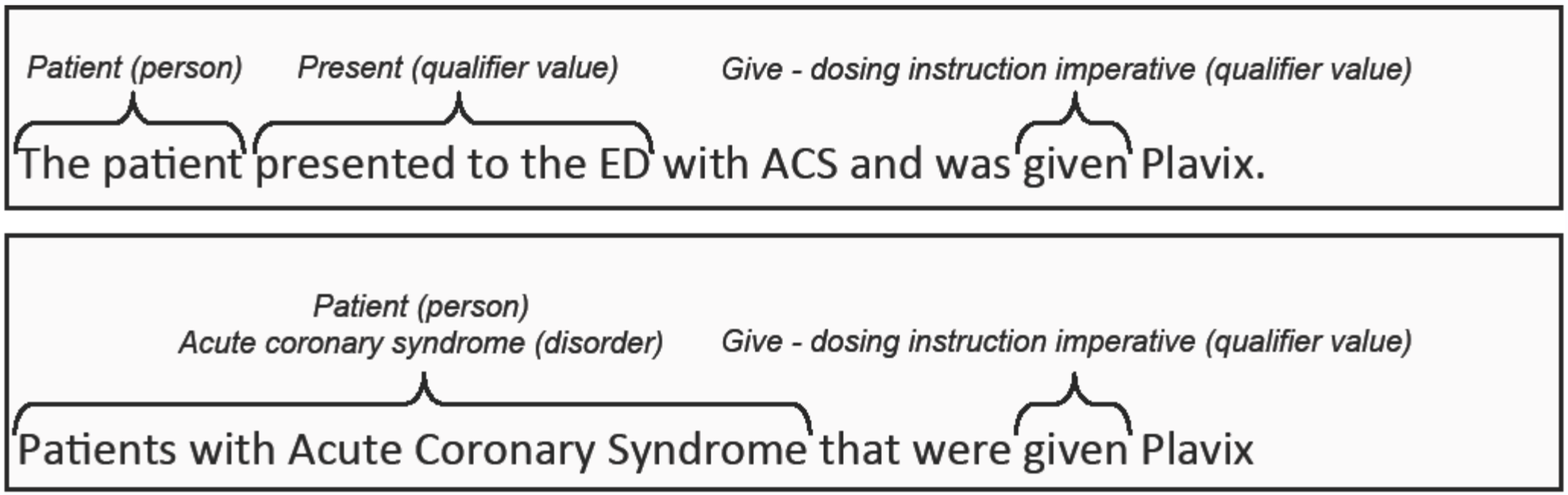
- **Semantic annotation**
  - The goal is to annotate free text with medical concepts and relationships from an ontology
  - Existing tools: MetaMap, cTakes
- **Semantic Search**
  - The goal is to retrieve documents that have been annotated with certain concepts
  - SPARQL

Limitations of semantic search



- **A concept might not exist in the ontology used in the annotations**
  - In the sample document, the drug Plavix does not exist in SNOMED CT and therefore no annotations are created for it
- **The annotation tools are not completely accurate**
  - The text fragment “presented to the ED” is annotated with the concept “Present (qualifier value)”, which is incorrect and incomplete

Limitations of semantic search



- **The annotations are limited**
  - Some tools do not annotate text with relationships
  - In the example, the fact that the patient was given Plavix is not explicitly represented in the annotations

Agenda

- Introduction
- Keyword search
- Semantic search
- **The ASE hybrid model**
- Conclusion

The ASE hybrid model

- **Based on two query languages**
  - Simple Ontology Query Language (SOQL)
  - Hybrid Query Language (HQL)
- **SOQL**
  - Subset of the SNOMED CT Query Language

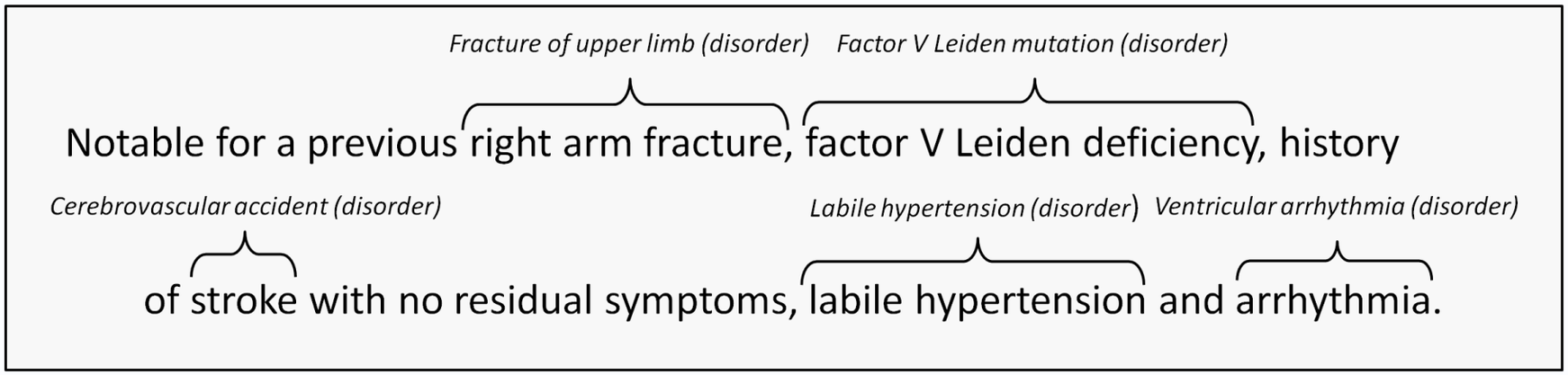
All	Parents(arg)
None	ParentsAndSelf(arg)
Self(conceptId)	Ancestors(arg)
Intersection (args)	AncestorsAndSelf(arg)
Union (args)	Descendants(arg)
Children (arg)	DescendantsAndSelf(arg)
ChildrenAndSelf(args)	HasRels(arg = arg, ...)

The ASE hybrid model

- **HQL: combines SOQL with free text**

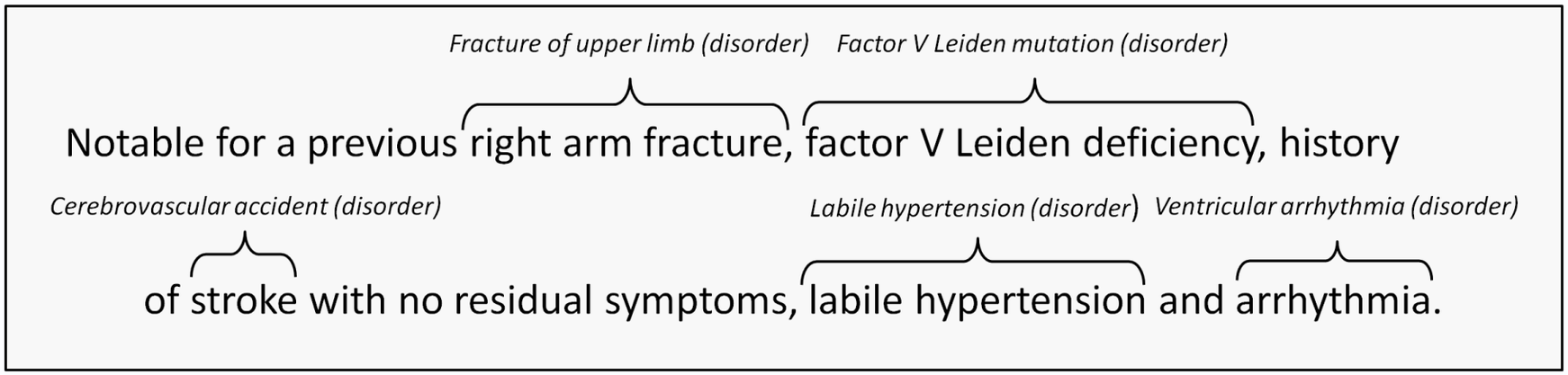
Query Type	Syntax
Term	*token*
Semantic	SOQL Expression
Boolean	(*sub-query, sub-query, ...*)
Annotated	< *sub-query, sub-query, ...*>
Ordered	*“sub-query, sub-query, ... ”*

Examples



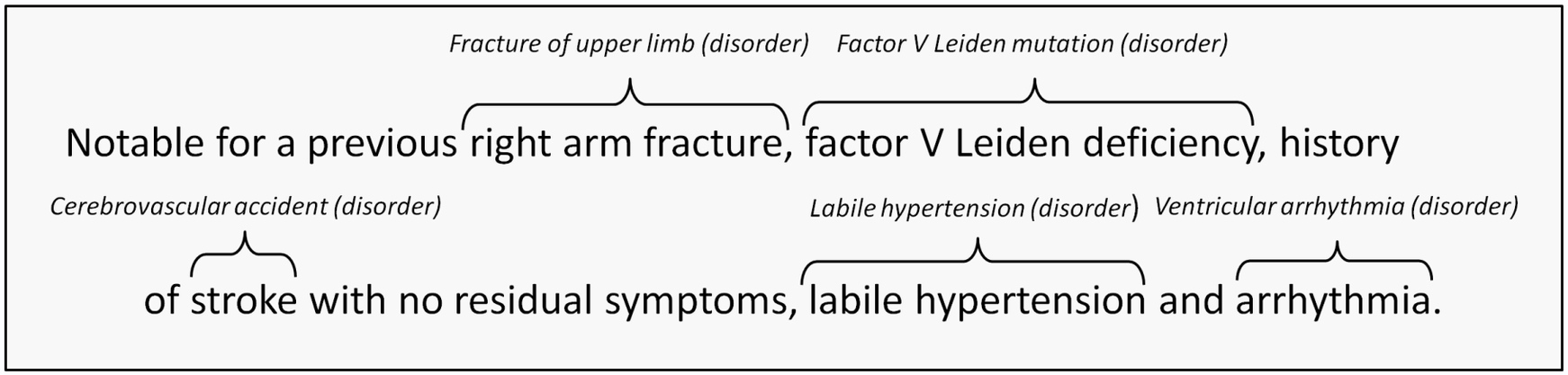
Query	Explanation
stroke	A term query returns documents that contain the specified keyword. This query would return the sample document.
Descendants(SCT_23853001 | Disorder of the central nervous system |)	A semantic query returns documents that contain any text annotated with the set of concepts that result from the evaluation of the embedded SOQL expression. This query would return the sample document because the SOQL expression evaluates to the set of all descendants of ‘Disorder of the central nervous system’, which includes ‘Cerebrovascular accident’.

Examples



Query	Explanation
(stroke appendicitis) *(+stroke +appendicitis)*	A boolean query evaluates each sub-query and returns either the intersection or the union of the matching documents. The first query will return the document because it contains the term ‘stroke’. The second query will **not** return the document because it also requires the term ‘appendicitis’ which is not present.
<Self(*SCT_230690007 |* *Cerebrovascular accident |)* *stroke>*	An annotation query evaluates each sub-query and returns the documents that contain terms with matching positions. This type of query is used to express annotation conditions. This query would return the sample document because it contains the term ‘stroke’ annotated with the concept ‘Cerebrovascular accident’.

Examples

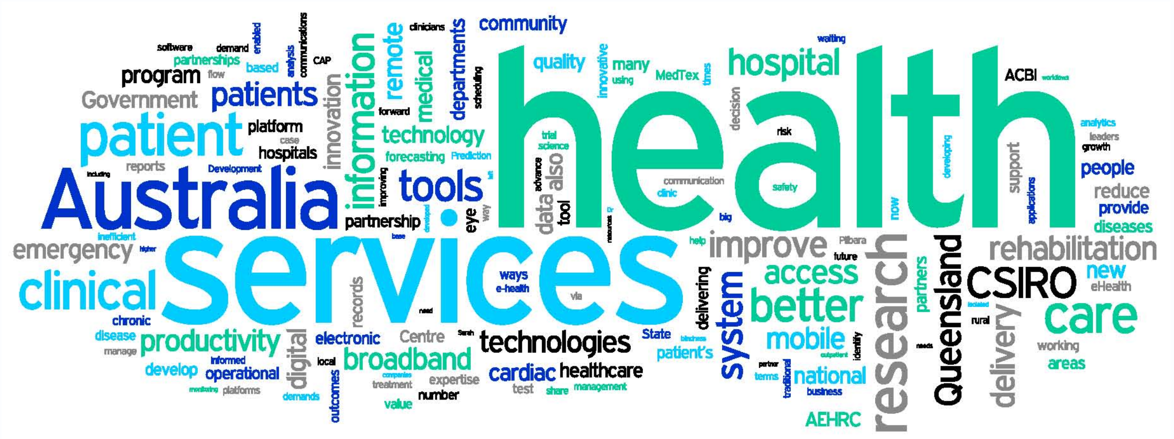


Query	Explanation
"previous DescendantsAndSelf(SCT_284003005 | Bone Injury |)"	An ordered query evaluates each sub-query and returns the documents that contain terms with adjacent positions in the specified order. The query will match the sample document because the first sub-query will return the document and the term token 'previous' with position {3} and the second sub-query will return the document and concept token 'Fracture of upper limb' (because a fracture of the upper limb is a specific type of bone injury) with positions {4,5,6}. Since the positions are adjacent and in the correct order, the document is considered a match.

Conclusion

- Existing tools can be used to annotated documents with medical concepts
- These annotations currently have several limitations
- The SOQL language can be used to search for documents annotated with medical concepts, but its effectiveness is affected by the quality of the annotations
- The ASE hybrid model allows combining semantic search with keyword-based search, allowing to overcome some of the limitations of semantic search

Questions?



Australian eHealth Research Centre

Michael Lawley — Group Leader
Alejandro Metke-Jimenez — Post Doc

Email: {michael.lawley, alejandro.metke}@csiro.au
Web: www.aehrc.com

THE AUSTRALIAN E-HEALTH RESEARCH CENTRE