TOWARDS SEMANTIC INTERPRETATION OF CLINICAL NARRATIVES WITH ONTOLOGY-BASED TEXT MINING

Bo Zhao
School of Computer Science and Informatics
Cardiff University

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD at Cardiff University.

2016

Abstract

In the realm of knee pathology, magnetic resonance imaging (MRI) has the advantage of visualising all structures within the knee joint, which makes it a valuable tool for increasing diagnostic accuracy and planning surgical treatments. Therefore, clinical narratives found in MRI reports convey valuable diagnostic information. A range of studies have proven the feasibility of natural language processing for information extraction from clinical narratives. However, no study focused specifically on MRI reports in relation to knee pathology, possibly due to the complexity of knee anatomy and a wide range of conditions that may be associated with different anatomical entities.

In this thesis, we describe KneeTex, an information extraction system that operates in this domain. As an ontology-driven information extraction system, KneeTex makes active use of an ontology to strongly guide and constrain text analysis. We used automatic term recognition to facilitate the development of a domain-specific ontology with sufficient detail and coverage for text mining applications. In combination with the ontology, high regularity of the sublanguage used in knee MRI reports allowed us to model its processing by a set of sophisticated lexico-semantic rules with minimal syntactic analysis. The main processing steps involve named entity recognition combined with coordination, enumeration, ambiguity and co-reference resolution, followed by text segmentation. Ontology-based semantic typing is then used to drive the template filling process. We adopted an existing ontology, TRAK (*Taxonomy for RehAbilitation of Knee conditions*), for use within KneeTex. The original TRAK ontology expanded from 1,292 concepts, 1,720 synonyms and 518 relationship instances to 1,621 concepts, 2,550 synonyms and 560 relationship instances. This provided KneeTex with a very fine-grained lexico-semantic knowledge base, which is highly attuned to the given sublanguage. Information extraction results were evaluated on a test set of 100 MRI reports. A gold standard consisted of 1,259 filled template records with the following slots: finding, finding qualifier, negation, certainty, anatomy and anatomy qualifier. KneeTex extracted information with precision of 98.00%, recall of 97.63% and F-measure of 97.81%, the values of which are in line with human-like performance.

To demonstrate the utility of formally structuring clinical narratives and possible applications in epidemiology, we describe an implementation of KneeBase, a web-based information retrieval system that supports complex searches over the results obtained via KneeTex. It is the structured nature of extracted information that allows queries that

encode not only search terms, but also relationships between them (e.g. between clinical findings and anatomical locations). This is of particular value for large-scale epidemiology studies based on qualitative evidence, whose main bottleneck involves manual inspection of many text documents.

The two systems presented in this dissertation, KneeTex and KneeBase, operate in a specific domain, but illustrate generic principles for rapid development of clinical text mining systems. The key enabler of such systems is the existence of an appropriate ontology. To tackle this issue, we proposed a strategy for ontology expansion, which proved effective in fast–tracking the development of our information extraction and retrieval systems.

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed …… (candidate) Date ……

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of ……(insert MCh, MD, MPhil, PhD etc, as appropriate)

Signed …… (candidate) Date ……

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.

Other sources are acknowledged by explicit references. The views expressed are my own.

Signed …… (candidate) Date ……

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed …… (candidate) Date ……

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans **after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.**

Signed …… (candidate) Date ……

Table of Contents

List of tables

List of figures

List of textboxes

Acknowledgement

I would like to express great thanks to my first supervisor, Dr. Irena Spasić, for her encouragement, never tired advice and guidance throughout my PhD. I would like to also thank Dr. Kate Button for her consistent professional knowledge support for this project.

Also special thanks to my second supervisor and panel meeting members, Prof. Chris Jones, Prof. David Marshall and Dr. Steven Schockaert. They have provided me with very useful and stimulating suggestions.

I would like to thank my dear parents for supporting and generously sponsoring my study.

I am also very grateful for the scholarship provided to me from Henry Lester Trust in 2012.

I would like to express my sincere appreciation to my supervisor, Dr. Irena Spasić, again, for helping proofreading and commenting this thesis.

Chapter 1 Introduction

1.1 Motivation

Electronic Health Record systems (EHRs) maintain patient medical history in digital format. Both US and EU have invested billions to promote the adoption of EHRs (Jensen et al., 2012). The adoption rate of EHRs among physicians had increased from 18% in 2001 to 57% in 2011 (Hsiao et al., 2010). And it is still continuously and rapidly increasing. The UK National Health Service (2015) has also planned to digitalise every GP practice by 2017.

Clinical narratives stored in EHRs normally describe a specific clinical event or situation. The expressiveness of natural language allows a physician to describe and share detailed situation of a patient with other professionals (NUANCE, 2008). Clinical narratives contain detailed information that includes but not limited to medications, diseases, observations, diagnostic processes and temporal information. it has been demonstrated that clinical narratives such as pathology and radiology reports could provide valuable information for both diagnostic and research purposes (Jensen et al., 2012; Mohanty et al., 2007; Spasic et al., 2005; Spasić et al., 2014).

Despite the great potential, such information is scattered and heterogeneous, which has limited direct analysis (Jensen et al., 2012). By applying natural language processing approaches such as text mining onto clinical narratives could help to improve this (Jensen et al., 2012; Meystre et al., 2008). Spasić et al. (2014) have also demonstrated the feasibility of using text mining to extract structured information in cancer related pathology and radiology reports.

Although the general text mining process could provide syntactic analysis to surface textual forms, real-world semantics connected to these surface textual forms is needed. Semantic interpretation on the other hand provides real-world interpretations that abstract meanings underneath their surface textual forms (Abbe et al., 2015; Albright et al., 2013).

Musculoskeletal conditions are the second largest contributor to years lived with disability (Vos et al., 2012). In the United Kingdom, a total of 33% of individuals aged 45 and over have sought treatment for osteoarthritis, with the knee being the most commonly affected joint (Arthritis Research UK, 2013). The incidence of acute knee injuries is reported to be at a rate of 2.29 per 1000 in US population (Gage et al., 2012). In the Netherlands, 45-55% of acute knee injuries develop into a long-term medical

condition (van Middelkoop et al., 2008). Patients may still experience movement deficiency 1 year following knee surgery (K. Button et al., 2014). In particular, participation restrictions may persist 2 years following total knee replacement (Maxwell et al., 2013).

If not diagnosed early and, therefore, not treated adequately, acute injuries can become chronic and lead to lifelong problems or conditions. When it comes to diagnosing knee pathology, magnetic resonance imaging (MRI) has the advantage of visualising all structures within the knee joint, i.e. both soft tissue and bone. When used in conjunction with medical history and physical examination, this makes MRI a valuable tool for increasing diagnostic accuracy and planning surgical treatments (Grover, 2012; Konan et al., 2009; Pompan, 2012; Wenham et al., 2014; Yan et al., 2011).

The MRI procedure is relatively simple and typically lasts 15 to 45 minutes. Following an MRI scan, the imaging results are summarised by a radiologist in a diagnostic narrative report that conveys a specialist interpretation of the MRI scan and relates it to the patient's signs and symptoms. Depending in the severity of their condition, some patients will need to be treated immediately while others may wait. An ability to interpret MRI reports automatically would help determine the priority of patients' treatments without the overhead associated with reading through a potentially large number of MRI reports. This can streamline the patient pathway and thus improve the prospects of their health outcomes.

1.2 Research question

A range of studies have proven the feasibility of natural language processing (NLP) for information extraction from clinical narratives. However, no study focused specifically using ontology-based text mining to support semantic interpretation on clinical narratives, possibly due to the complexity of knee anatomy and a wide range of conditions that may be associated with different anatomical entities. Using this specific domain as a case study, we wanted to test the following hypotheses:

- Semantic interpretation of clinical narratives can be automated with ontology-based text mining.
- Ontology expansion can be effectively achieved using a systematic pipeline.

Automatic semantic interpretation requires machine-readable knowledge representation. Ontologies are commonly used to formally describe domain-specific knowledge and facilitate information exchange. Ontologies can be coupled with NLP techniques (such

as information retrieval and information extraction) to facilitate navigation through large volumes of text data. The "big data" aspect is of particular importance for epidemiology, the study of the distribution and determinants of health-related states/events in defined populations. By its definition, epidemiology relies heavily on statistical methods. Unfortunately, many of published research findings are probably limited due to sampling bias and low statistical power. Given the complexity and cost of manual interpretation of clinical narratives, it is not surprising that the size of such epidemiologic studies has been limited to hundreds or even dozens of cases. If interpretation of evidence described in clinical narratives such as those found in MRI reports could be automated, then it would overcome the size limitation in retrospective cohort studies posed by the need to manually sort through the evidence.

The idea of combining ontologies and NLP to support natural language understanding is by no means new. However, the extent of knowledge engineering involved in the development of domain–specific ontologies with sufficient detail and coverage for NLP applications is known to present a major bottleneck in deep semantic NLP. One of the contributions of this thesis is an approach that can be used to rapidly develop ontologies that can act as very fine–grained lexico–semantic knowledge bases that are highly attuned to the targeted sublanguage.

1.3 Research outcomes

1) Rapid ontology development framework

We adopted an alternative approach based on a set of strategies, primarily data-driven, which can be used to systematically expand the coverage of existing ontologies or to develop them from scratch. Three of these strategies are data–driven and as such are more likely to ensure that the ontology effectively supports the intended NLP application. Each data–driven strategy utilises a different approach to extracting the relevant terminology from the data either manually or automatically. The fourth strategy is based on integration of concepts from other relevant knowledge sources. The two main aims of this strategy are: (1) to avoid the over-fitting of the ontology to limited data available, and (2) to provide an initial taxonomic structure to incorporate new concepts.

We also illustrated how these strategies were implemented in practice to expand the coverage of the TRAK ontology, which is an ontology developed for modelling information related with rehabilitation of knee conditions, to make it suitable for a specific NLP application.

2) Expanded TRAK ontology

We practically demonstrated the approach to update the TRAK ontology in order to allow interpretation of information contained in knee MRI reports. We expanded TRAK into a fine-grained ontology from 1,292 concepts, 1,720 synonyms and 518 relationship instances to 1,621 concepts, 2,550 synonyms and 560 relationship instances.

3) KneeTex

We developed KneeTex as a human-level performance ontology–driven system for information extraction from narrative reports that describe an MRI scan of the knee.

As an ontology–driven information extraction system, KneeTex makes active use of the TRAK ontology to strongly guide and constrain text analysis. The main processing steps involve named entity recognition combined with coordination, enumeration, ambiguity and co–reference resolution, followed by text segmentation. Ontology–based semantic typing is then used to drive the template filling process. On a gold standard, KneeTex extracted information with precision of 98.00%, recall of 97.63% and F–measure of 97.81%, the values of which are in line with human–like performance. These results confirmed that, when having an appropriate ontology, we can fully automate semantic interpretation of facts about pre–specified types of entities and relationships from knee MRI reports.

4) KneeBase

We demonstrated KneeBase as an example of the integrated use of ontology and extracted information to build an information retrieval system. The system structure and programming framework can be reused for other similar ontology-based information retrieval tasks.

The basic principles of all these contributions are generic and are applicable in other domains.

The main results reported in this thesis have been published in the following article:

> Irena Spasić, Bo Zhao, Christopher Jones, Kate Button (2015) *KneeTex: An ontology-driven system for information extraction from MRI reports*. Journal of Biomedical Semantics, Vol. 6, 34

1.4 Thesis structure

Chapters 2 and 3 provide a literature review. Chapter 2 provides a review of natural language processing, specifically in relation to its clinical applications. Chapter 3 deals with clinical knowledge representation and focuses specifically on ontologies. Chapters 4 to 7 represent the original contribution of this thesis. Chapter 4 describes the data used in this study. It is used to illustrate the nature and complexity of semantic interpretation of knee MRI reports in addition to setting a gold standard for its evaluation. Chapter 5 describes our approach to efficiently expanding the TRAK ontology in order to make it fit for interpreting information contained in knee MRI reports. Chapter 6 describes the KneeTex system specification, including its specification, specific design and implementation choices and evaluation. Chapter 7 describes a separate information retrieval system structure, which demonstrates a potential of using KneeTex to support epidemiological research. Finally, we conclude the thesis by highlighting the main contributions to knowledge engineering and NLP, and suggest possible directions of future work.

Chapter 2 Natural language processing

Natural language processing (NLP) is a subset of computational linguistics. NLP can be defined as the automatic or semi-automatic processing of human language from aspects such as morphology, syntax, semantics and pragmatics (Copestake, 2004). NLP includes many subareas, such as parsing, knowledge representation, ontology, information extraction, text mining, etc.

NLP is now used for analysis in many domains. However, a well performed NLP application in a specific domain may not achieve the same performance in other domains due to different language characteristics among different domains as well as biased annotated corpus. Therefore, when switching domain, a substantial level of performance degradation is usually expected (Cambria and B. White, 2014; J. Jiang, 2008; L Sumathy and Chidambaram, 2013).

The general computational linguistics and NLP terminologies and theories have become an obstacle for researchers from other domains due to different criteria, specific knowledge and restricted sublanguages available in different domains. As a result, specific domain related NLP processes have been carried out to enable practical applications (Wintner, 2009).

In the earlier days, most medical-related data was stored in the form of paper instead of stored digitally nowadays, e.g. medical records, patient notes handwritten by doctors or nurses, prescriptions, X-ray report, MRI report, etc. With the development of digital storage, computing platforms and Internet, medical data is now rapidly transforming to digital format. This enables researchers easier access to medical data and the possibility to carry out further research without converting paper based data to digital format.

This chapter introduces the concept of sublanguage and its characteristics, especially focusing on clinical sublanguage. Typical information extraction and named entity recognition processes and achievements are also described. To help understanding NLP applications in the clinical domain, we have also provided a review on state-of-the-art clinical NLP systems.

2.1 Sublanguage

Languages are made up of many different word sequences with certain restrictions to express meanings. These restrictions include syntaxes, co-occurrence patterns, and lexical restrictions (Harris, 1991).

A practical application of NLP often involves a specific domain. This specific domain is also called restricted domain in NLP concepts. Sublanguage is the language used in a restricted domain, which is normally used for communications among domain experts (Starren and S. M. Johnson, 1996). The development of a sublanguage normally comes through the use of general language in under a specific circumstance. Therefore, a sublanguage usually has its own characteristics that distinguish it from the general language as well as some restrictions inherited from the general language (Harris, 1991; Lehrberger, 2014; Zeng et al., 2011).

2.1.1 **Characteristics**

Although sublanguage inherits some restrictions from its general language, it may develop some lexical, syntactic, semantic, terminology and sentence structures that cannot be applied back to its general language (Zeng et al., 2011). Characteristics of sublanguage normally include:

- Semantic classifications

  Semantic classifications can be applied to relevant words in a sublanguage. Names of classified categories are usually informational and can be words from the sublanguage, such as *disease*, *injury*, *body part*.
- Co-occurrence patterns

  A sublanguage is usually structured by frequently repeated co-occurring patterns to form meaningful relations, such as *bone marrow edema* and *tear in medial meniscus*. These patterns may not have different meaning if seen in general language or other sublanguages. However, the frequencies of those co-occurrences can be one of its characteristics that differ from other sublanguages (Sager et al., 1994). It is also possible to have restrictions to constrain which types of words can occur together (Friedman et al., 2002).
- Paraphrastic transformations

  Different expressions can be adopted to represent the same meaning. This often involves changing of part-of-speech and tense of words (Friedman et al., 2002; Hirschman and Sager, 2002). For example, the following representations are all equivalent to each other.
  - `SYMPTOM + ADJECTIVE + BODYPART`: *tear at lateral meniscus*
  - `SYMPTOM + BODYPART`: *torn lateral meniscus*
  - `BODYPART + VERB`$_{BE}$ `+ SYMPTOM`: *lateral meniscus is torn*

- Contextual omission

  With given context for a specific domain, implicit information can often be uncovered with sufficient knowledge of the context (Friedman, 2006; Friedman et al., 2002; Hirschman and Sager, 2002). For example, in a MRI report of the knee, *normal extensor tendons* would be interpreted as *normal quadriceps tendon and patellar tendon* (Sonin et al., 1995).

- Terminology

  Vocabulary in a sublanguage is restricted. Only limited words of certain classifications and limited co-occurrence patterns are allowed. These terms may also represent connected but different meanings other than in the general language. The frequency of occurrences of these terms are also expected to be much higher than in the general language (Ciravegna, 1995; Friedman, 2006; Friedman et al., 2002; Leroy and H. Chen, 2001; Nkwenti-Azeh, 2001), such as *capsule* in knee MRI reports as anatomy structure surrounding the joint and *capsule* in general language as a small container.

2.1.2 **Clinical sublanguage**

Clinical language has a long developing history since the Greek era. It has always come with great challenges for linguists (Wulff, 2004). Most clinical professionals tend to use clinical language more often other than using general language when communicating with other professionals (Bourhis et al., 1989). As soon as Harris introduced the term *sublanguage*, language used by clinicians was recognised as a proper sublanguage (Patterson and Hurdle, 2011). However, clinical language can still be divided into many sublanguages with partly overlapped coverages as there exist sub-domains of the overall clinical domain (Friedman, 2000; Stetson et al., 2002).

Clinical language also has the same characteristics as discussed above, i.e. semantic classifications, co-occurrence patterns, paraphrastic transformations, contextual omission and terminology. Campbell and Johnson pointed out that there are significant differences between clinical and non-clinical texts (Campbell and S. B. Johnson, 2001). Despite of frequently used terminologies, clinical texts tend to use some part-of-speech more often than general English, as well as some co-occurrence patterns, see ***Table 2-1*** and ***Table 2-2***. It is very clear that there are significant higher level usages of nouns, adjectives and numbers in clinical texts, which is reasonable because clinical texts primarily consists of clinical findings, anatomy parts and severity descriptions (Friedman et al., 2002).

As a language for professional communication, clinical language tends to be very concise (Stetson et al., 2002). Simple noun phrases or short sentence fragments are favoured over complete and complex sentence structures. Therefore, abbreviations are frequently used instead of their complete forms, such as *ACL* for *anterior cruciate ligament*, *MCL* for *medial collateral ligament*, and *NAD* for *no active disease*.

Table 2-1 Part-of-speech frequency distributions in clinical and non-clinical english texts (Campbell and S. B. Johnson, 2001)

Part-of-speech	Clinical	Non-clinical
NUMBER	5319	1989
$VERB_{base\_form}$	1260	4345
$NOUN_{singular}$	22615	17389
$PROPER\_NOUN_{singular}$	2740	5755
$VERB_{past\_tense}$	6613	4171
MODAL	203	1188
ADJECTIVE	11383	8928

Table 2-2 Part-of-speech bigram frequency distributions in clinical and non-clinical english texts (Campbell and S. B. Johnson, 2001)

Part-of-speech Bigram	Clinical	Non-clinical
$NOUN_{singular}$ + $VERB_{past\_tense}$	3231	1084
$PROPER\_NOUN_{singular}$ + $PROPER\_NOUN_{singular}$	599	2572
$VERB_{past\_tense}$ + $VERB_{past\_participle}$	2014	470
NUMBER + NUMBER	1255	216
$NOUN_{singular}$ + ADJECTIVE	1340	339
$VERB_{past\_tense}$ + ADJECTIVE	1410	381

2.1.3 **Challenges in clinical NLP development**

Substantial problems exist when NLP development moves from general language to the clinical domain. However, these problems are not restricted to the clinical domain only. They could apply to NLP development towards any high skilled restricted domain (Chapman et al., 2011). These problems present in many aspects such as access to clinical data, availability of annotated datasets, availability of sufficient domain knowledge and availability of standard formats.

Unlike developing NLP applications for general language, clinical NLP development requires possession of relevant domain knowledge. Lack of such knowledge would lead to difficulties in interpreting implicit information underneath partly omitted expressions and abbreviations. Lack of domain knowledge also reflects in the availability of properly annotated datasets. Although clinical sublanguages have partly overlapped coverages, they still differ in syntaxes and semantics. A properly annotated dataset is the premise for training and evaluation processes (Chapman et al., 2011; Friedman et al., 2002).

The access to clinical text is also restricted by certain laws and rules protecting patient privacies. For example, in the UK, access to patient records are governed by many NHS Code of Practices and laws, including the The Data Protection Act 1998 (UK Department of Health, 2003).

Clinical text contains many abbreviations that are used for convenience. These abbreviations are often derived from the record writer itself based on its own domain knowledge and some unregulated common conventions (Raja and Jonnalagadda, 2015). Such abbreviations would lead to a high level of ambiguities for NLP processes. For example, *PT* in clinical text have many interpretations such as *patient* and *physical therapy*, etc. Similar with the abbreviation problem, synonyms are also being used frequently. For example, *rupture*, *disruption* and *split* all point to the concept of *tear*.

Although there exist many clinical text templates, in most cases, clinical text still uses the format of free text, which allows doctors to write precise notes with the flexible expressions (van Ginneken et al., 1997). A lot of elements in clinical text that can be helpful for clinical NLP developments are often missing inappropriately used, such as section headings, punctuations (Raja and Jonnalagadda, 2015).

Differences in co-occurrence patterns and semantic classifications among clinical sub-domains also lead to expected performance decreases if a clinical NLP system developed for one sub-domain being applied to another (Friedman, 2000; Patterson and Hurdle, 2011). Hence, the portability of a clinical NLP system is quite limited.

2.2 Information extraction

The aim of information extraction is to recognize meaningful information from free text. Such meaningful information normally represents some real-world entities, e.g. event, findings, locations and relevant attributes, etc. Instead of providing information consumers with entire text, these meaningful elements will be picked out and presented. Extracted information will be in a structured format and provide the ability to be directly stored in databases for query and further analysis (Piskorski and Yangarber, 2013).

Information extraction originates from template-filling tasks of the Message Understanding Conference (MUC) (Moens, 2006). Templates are final output results for information extraction task. A template is a structured representation of extracted information. An example of template for medical records will be similar as shown in ***Figure 2-1***.

In each template, there are several slots. For example, in ***Figure 2-1***, the template including following slots: patient name, gender, admission date, age, record date, symptoms, department and treatments. These slots need to be filled with relevant extracted information.

Medical Records Template	
PATIENT NAME:	GENDER:
ADMISSION DATE:	AGE:
RECORD DATE:	SYMPTOMS:
DEPARTMENT:	TREATMENTS:

Figure 2-1 A template for medical records information extraction

Information that can be extracted and filled into these slots normally shows up in free texts frequently with consistent co-occurrence patterns. These patterns can be used to help identifying and extracting relevant information. Template designs and patterns can vary a lot for different domain related information. In each domain, it may contain several scenarios, which are some specific events or relations. A scenario can be surgery information, pharmacotherapy information or else in the domain of medical notes (Grishman, 1997; Muslea, 1999).

Early stage information extraction systems normally based on rules and patterns recognised by human experts through iterating corpus text. Manual approaches could provide sufficient precision but lack extensibility. It is normally limited to a specific scenario, domain and language. And the whole process is time-consuming. Later it started to develop towards the direction of solving general purpose tasks and extendible system frameworks.

To reduce human burdens in information extraction systems, machine learning approaches had been gradually introduced. Although initially limited to supervised learning, it still started leading the trend of moving from traditional knowledge engineering approaches to trainable approaches. Traditional knowledge engineering approach requires input from both system designer and domain experts. But with trainable machine learning approaches, it only requires domain experts input on training corpus annotation when switching to a new scenario, domain or language as the system structure can be reused.

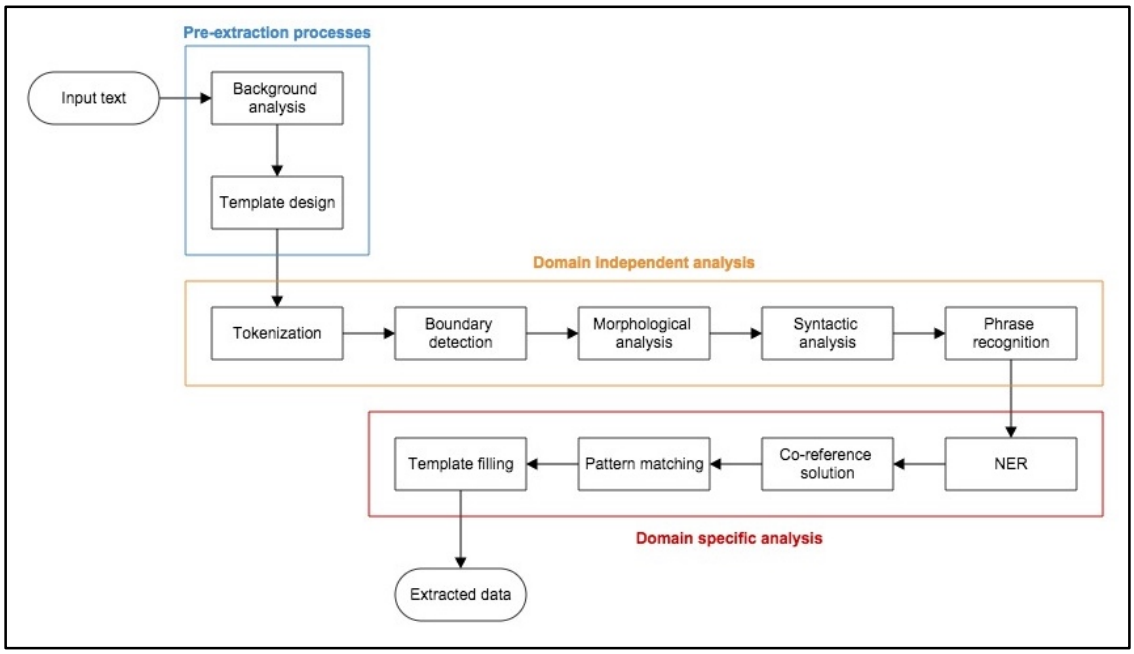


Figure 2-2 A common information extraction system structure (Piskorski and Yangarber, 2013)

Information extraction systems may have different structures for different tasks. However, there are some common components in most information extraction systems. ***Figure 2-2*** shows a common information extraction system structure (Piskorski and Yangarber, 2013):

a) Background analysis collects information about the whole input dataset, including input format, data structure, related domain and relevant metadata.
b) With domain information and data structure obtained from background analysis, template format and slots can be determined.
c) Tokenization split texts into word-like tokens. Meanwhile, POS tagging will be performed to classify token types.
d) Boundary detection helps to segment input text into sentences.
e) Sometimes a word can have varied forms depending on its tense, person, etc. Morphological analysis could use POS tags and lemmas to help with disambiguation.
f) Phrase recognition extracts small structures from input texts. These small structures are normally common phrases or frequently co-occurred meaningful words.
g) Named entity recognition is performed to identify entities from input texts, both domain independent and domain specific. Domain independent entities include

temporal information, numbers, currencies, etc. Domain specific entities will vary as focus domain changes.

h) With all previous processes, patterns can be recognized or compiled to identify and extract relevant concepts to fill the slots in previous designed template.

i) Co-references happen frequently in natural language. Demonstrative pronouns, such as *this*, *that*, *it*, etc., are quite often used to reference an entity mentioned in previous text. Co-references solutions will help to identity which entity the demonstrative pronoun refers to.

Template filling stage will allocate previous extracted entities into slots in predefined template.

2.2.1 **Named entity recognition**

The concept of named entity (NER) was proposed during the 6th Message Understanding Conference (MUC-6). It was one of the subtasks that extracts person, location and organization related terms using tag ENAMEX, short for entity name expression (Grishman and Sundheim, 1996). The process of extracting these named entities is called NER, short for named entity recognition.

Named entity is now used to identify basic elements in text that can be classified into pre-defined categories. These categories may include *persons*, *locations*, *organizations*, *time* and some other domain related categories. For example, in this knee injury research, there are categories like anatomical locations, findings, etc.

Named entity recognition approaches in different domains may vary. Most NER systems are specifically dedicated to certain domains, e.g. emails, gene information, etc. Although there may be some general rules or patterns that apply to most domains, extending a current NER system into another domain without making proper tuning will lead to significant drop in its performance (Nadeau and Sekine, 2007). Meanwhile, named entity tags can be different for different domains as different domain sublanguages may use different semantic classifications as mentioned in ***2.1.1***. For example, in clinical domains, the tag *locations* may be replaced by *anatomical entities*.

Similar to POS tagging, NER can be performed using either manually rule-based or machine learning approach. Manually compiled rules often rely on grammatical, syntactic and morphological features, as well as dictionaries if possible. See ***Textbox 2-1*** for a basic dictionary-based NER rule. In this example, a title *Dr.*, and a name *Banner* are recognised

using predefined dictionary. Based on the rule $RULE_{PERSON}$: TITLE + NAME, *Dr. Banner* is recognised as a person entity.

```
DICTIONARY_TITLE : Mr.,Mrs., Ms., Dr., Prof., ...
DICTIONARY_NAME : Abigale, Banner, Brown, Colin, ...
RULE_PERSON: TITLE + NAME
EXAMPLE:
<title>Dr.</title> <name>Banner</name> found an acute grade III MCL injury.
```

Textbox 2-1 A basic NER manual rule example

Although manually compiled rules have been proven that they have better performance than machine learning approaches in restricted domains and complex entities, compiling these rules manually can be time consuming, tedious and error prone. Disadvantages of manually compiled rules are also obvious. Performance will drop significantly if a manually compiled rule was applied to another domain, as manually compiled rules are weak on portability. Even a slight change in the same domain, compiled rules might need to be rewritten. Therefore, manually compiled rules are often restricted by domain and language (Mansouri et al., 2008).

Machine learning approaches use classification models to perform identification process through supervised or unsupervised methods. At least one training dataset is required for supervised learning. The training dataset normally contains pairs of input and output, of which the correct output is usually annotated and curated manually. This paired dataset together with given set of parameters are then used to generalise a relationship model between inputs and outputs (Lison, 2012; Maimon and Rokach, 2005). Supervised NER can be achieved using various statistical models, including hidden Markov model, maximum entropy, decision trees and support vector machines, etc (Mansouri et al., 2008).

Unsupervised learning methods are normally used when available training dataset contains only input information. Pure unsupervised approaches are quite unusual. A pure unsupervised NER system will learn from given data without any human input. Normally unsupervised approaches are used to refer to weakly supervised approaches or semi-supervised approaches. A weakly supervised approach usually includes human error correction stage where automatically induced patterns will be reviewed and corrected by human (Hastie et al., 2008; Maimon and Rokach, 2005; Mansouri et al., 2008).

Rule-based approaches and machine learning approaches could also be combined together to make hybrid approaches. Hybrid approaches will combine advantages from both rule-based approaches and machine learning approaches. Hybrid approaches normally will have better performances. As hybrid approaches still include rule-based part, it still requires a stable restricted domain (Mansouri et al., 2008).

2.2.2 **Clinical named entity recognition**

Clinical named entity recognition, sometimes also known as clinical concept extraction, is an important aspect of clinical natural language processing. Extracted clinical entities, such as symptoms, diagnosis and treatments, form the basis to interpret underlying knowledge and support further research (Jonnalagadda et al., 2012; Tang et al., 2013).

Clinical NER has been significant attentions in recent years. Following the MUC conference for information extraction (Grishman and Sundheim, 1996), The Informatics for Integrating Biology and the Bedside (i2b2) has been holding annual challenges focusing on the clinical domain since 2006 (Uzuner et al., 2006). ***Table 2-3*** lists out entities that have been focused on in the past i2b2 challenges.

i2b2	Targeting entities
2006	Private health information (PHI)
2006	Smoking status
2008	Obesity
2009	Medication
2012	Temporal information
2014	Heart disease risk factors
2014	Private health information (PHI)

Table 2-3 Yearly i2b2 shared task challenges (Uzuner, 2009; Uzuner et al., 2006; 2008; 2007; 2010; 2011; Uzuner and Stubbs, 2015)

In general, systems that integrated statistical learning and regular expression rules have the best performances in identifying PHI in i2b2 2006 (Uzuner et al., 2007). Wellner et al. (2007) achieved the best F-measure of 0.9736 for this challenge using their submission built on Conditional Random Fields (CRFs) based Carafe system and regular expression patterns.

Clark et al. (2008) built a system that combines Support Vector Machine (Vapnik, 1982) based supervised learning and linguistic rule-based extraction. This system achieved the best micro-averaged F-measure of 0.90 compared with other submissions. Clark et al. (2008) suggested that the use of linguistic rules played an important part in such achievement, which helped reducing limitations caused by restricted training set size.

In 2008 i2b2 challenge of identifying obesity entities (Uzuner, 2009), the challenge is divided into two subgroups: the textual task for identifying explicitly described obesity terms and the intuitive task for identifying descriptions implicitly describing obesity. Rule-based approaches and machine learning approaches contributed most to the textual task and intuitive task respectively. Rule-based systems manually curated by domain experts proved to be very efficient in this closed-domain obesity challenge. However, generalisability is an issue for these systems if were to be applied to new tasks and domains.

The 2014 i2b2 challenges of identifying heart disease risk factors have overlapped with some of previous challenges on smoking status, obesity, medication and temporal information. Therefore, it benefits from previous tasks for have an enriched annotated training set. This has proved to have the biggest positive affect on the success of submitted systems (Stubbs et al., 2015). For example, the best performance on past smokers category has an increased F-measure of 0.8869 compared with 0.67 in 2006. The U.S. National Library of Medicine (NLM) (Roberts et al., 2015) team achieved the best performance in i2b2 2014 is a micro-averaged F-measure of 0.9276. The team devoted large effort on re-annotation of 2/3 of the training corpus, which has greatly improved the system performance from 0.9021 to 0.9276 in F-measure.

2.2.3 **Evaluation**

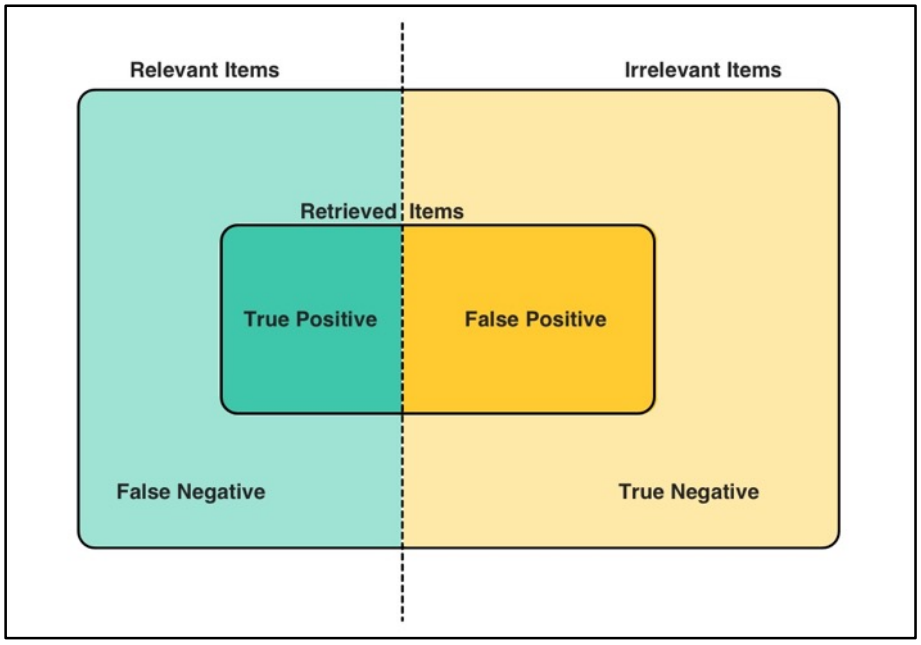


Figure 2-3 Introduction of precision and recall (Maedche, 2012)

Precision, recall and F-measure are commonly used measures in information extraction evaluation. These three measures are originally used in information retrieval systems. ***Figure 2-3*** gives a brief idea of precision and recall. Precision measures the amount of retrieved relevant items and recall measures relevant items that are retrieved.

$$Precision = \frac{True\ Positive}{True\ Positive\ + False\ Positive}$$

$$Recall\ = \frac{True\ Positive}{True\ Positive\ + False\ Negative}$$

Definitions:

- True Positive: number of retrieved items that belongs to relevant items
- True Negative: number of irrelevant items that are not retrieved
- False Positive: number of retrieved items that are irrelevant
- False Negative: number of relevant items that are not retrieved

Precision measures how many items are correctly retrieved. Recall measures how many items that should be retrieved are actually retrieved. Normally, we would prefer precision and recall both as high as possible at the same time. However, sometimes system precision and recall values are conflicted. Therefore, F-measure, which is a weighted average of precision and recall, is introduced and also often used as a key measure (Hripcsak and Rothschild, 2005).

$$F = \frac{(1 + \beta^2) * precision * recall}{(\beta^2 * precision) + recall}$$

β is the weighting of precision towards recall. Sometimes if overlapped matching is allowed when calculating precision and recall, there would be different weighting value. Normally precision and recall are treated equally (Maynard et al., 2006). In this case, β is set to 1, so the equation can be simplified to:

$$F = \frac{2 * precision * recall}{precision + recall}$$

2.3 Clinical NLP applications

Clinical NLP applications are normally developed to solve specific questions. There are many questions in the clinical domain, such as identification of risk factors for a specific disease (Uzuner and Stubbs, 2015), identification of medical records from documents. Meanwhile, there have also been many systems available already to solve them.

2.3.1 **MetaMap**

MetaMap is a concept search program built based on the UMLS Metathesaurus. It maps text input with UMLS Metathesaurus and other relevant knowledge bases to identify biomedical concepts mentioned in text (Aronson, 2001).

The original MetaMap program was developed to improve the retrieval of bibliographic materials such as MEDLINE citations.

MetaMap uses two parsers together to parse input text: the SPECIALIST minimal commitment parser (McCray et al., 1994) and the Xerox POS tagger (Cutting et al., 1992). The SPECIALIST parser identifies phrases, and allocates SPECIALIST lexicon tags. There are 11 SPECIALIST lexicon tags: *noun*, *verb*, *adjective*, *adverb*, *auxiliary*, *modal*, *pronoun*, *preposition*, *conjunction*, *complementizer* and *determiner*. If there would be any word that does belong to the lexicon, it will then be passed on to the Xerox POS tagger for normal POS tagging. A group of variants will be generated for each phrase that can be recognised using SPECIALIST lexicon and the supplement synonym database. These variants include acronyms, abbreviations, synonyms, derivational variants, inflectional variants and spelling variants with edit distance score (Aronson, 1996). With generated variants for each phrase, a set of candidates will then be retrieved from the Metathesaurus. Various optional parameters could be added to this process for better efficiency and precision, such as occurrence frequency, stop words and user-defined indexes. Each candidate will be evaluated towards matched phrase from the input text and provided a candidate score from 0 indicating no match at all to 1000 indicating perfect match (Aronson, 2006). Candidates will then be joined together to make mapping suggestions. Mapping suggestions also come with mapping scores where higher score indicates more confident MetaMap interpretation of matched text.

With no further implementation, MetaMap achieved a precision range from 53.19% to 91.30%, and a recall range from 94.59% to 100% in a recent evaluation towards a set of radiographic image reports (Al-Safadi et al., 2013). However, there were also a few weaknesses spotted:

- The Metathesaurus that MetaMap was built on does not cover every concept. As a result, uncovered concepts cannot be retrieved using MetaMap.
- Relationships within the phrase are ignored. For example, MetaMap output for *a tear found in lateral meniscus* will be two individual concepts *tear* and *lateral meniscus*.

- Word variants with punctuations such as dots and hyphens cannot be retrieved, e.g. *M.R.I* and *cardio-thoracic*.
- Concept names and variants in the Metathesaurus sometimes are different from commonly used terms in clinical free text.
- MetaMap does not handle spelling mistakes very well.

2.3.2 **MEDLEE**

MEDLEE, short for *Medical Language Extraction and Encoding System*, is a NLP system originally developed to extract clinical information from patient reports and store into a database for other programs to use (Friedman et al., 1994). It analyses input text both semantically and syntactically. It attempts to analyse the whole sentence as a start with additional analysis on segmented phrases if the first attempt failed (Friedman et al., 2008).

MEDLEE turns input patient radiology report into structured data through four steps: pre-processing, parsing, phrase recognition and encoding (Friedman et al., 2008).

Pre-processing Unformatted and formatted information is separated and extracted from the input patient report. A lexicon is used to retrieve specialized expressions used in relevant sublanguages to ensure the accuracy of retrieved information. A sentence will be saved for further parsing process if it contains any word not included in the lexicon.

Parsing Medical sublanguage expressions for the same concept may have quite different syntactic structures, but will normally share the same semantic structure. Therefore, the parser uses a grammar that focuses on semantic structures, but also includes some syntactic features. Variants that cannot match the lexicon but match predefined semantic grammar, will be sent for phrase recognition. However, if a sentence that its components match neither any lexicon term nor the semantic grammar, it will not be processed.

Phrase recognition Sometimes a multi-word may occur in a discontinuous form, e.g. *enlarged heart* may be separated by other words and become *heart appears to be enlarged*. This component uses parsing output to compare with a multi-word phrase mapping database to identify these phrases.

Encoding To keep consistency with its controlled vocabulary, retrieved information will then be compared to a synonym database to match controlled vocabulary terms for the final output.

MEDLEE was evaluated on a set of randomly selected 230 radiology reports, focusing on four types of diseases: neoplasm, chronic obstructive pulmonary disease, acute

bacterial pneumonia and congestive heart failure. Three physicians were involved in this evaluation. They had also helped creating the gold standard. MEDLEE achieved an average precision of 87% and an average recall of 70% (Friedman et al., 1994). However, it is found that although MEDLEE has good performance in precision, its performance in recall is relatively lower than other tools (Barrows et al., 2000).

2.3.3 **MPLUS**

MPLUS is a probabilistic model based medical text process tool. Different from other rule-based tools, MPLUS uses a semantic model built on Bayesian Networks to enable the capability of inferring patterns from contexts as well as with predefined rules (Christensen et al., 2002).

The semantic Bayesian Networks model used in MPLUS not only provides a semantic structure, but also works as a probabilistic inference engine. Unlike other NLP tools that perform semantic analysis and syntactic analysis separately, MPLUS integrates the semantic Bayesian Networks into its syntactic analysis process. In other words, the semantic analysis and syntactic analysis in MPLUS are interleaved and mutually constrained. Each recognised word-level phrase will be created an instance in the semantic Bayesian Networks model. These instances can also be united to match larger grammatical patterns. A probability value is also assigned to each model instance. Probability values of phrases are determined by its associated instances. Phrases are processed in the order of their probability. This process is mutually constrained syntactically and semantically. If a phrase with correct grammar cannot be semantically interpreted, it will be rejected. Similarly, if a interpretable phrase has a low probability value, it is also very unlikely that it will be included in the final output (Christensen et al., 2002).

MPLUS was evaluated on a set of 2,600 head CT (computerized tomography X-ray) scan reports, of which 600 were randomly selected to form the test set and the rest 2,000 reports become the training set. The task is to classify these reports into positive and negative groups, of which positive means present of suspected diagnosis concepts. A gold standard is created by four certified physicians. An average recall of 88% and average precision of 86% is achieved in the gold standard. MPLUS achieved an average recall of 87% and an average precision of 85%, which were both quite close to the gold standard performance (Fiszman et al., 2002). A pure rule-based and knowledge-based NLP system may be incomplete while applying it to a specific domain. The use of semantic Bayesian

Networks model allows MPLUS to process partially matched terms and terms that are not included in associated knowledge bases.

2.3.4 **cTAKES**

cTAKES, short for *clinical Text Analysis and Knowledge Extraction System*, is an open source NLP system designed to extract information from free-text clinical reports based on UIMA and OpenNLP (Savova et al., 2010).

cTAKES processes input text using a module-based pipelined system structure. It divides input text into sentences using punctuations such as period, question mark, and exclamation mark. Tokenization uses a 2-step method: sentences are split into tokens using spaces and punctuations; relevant tokens are rejoined together to create specific types of tokens such as date, time, measurement, and personal titles. Tokens are then normalized to match the lexicon to improve the tolerance of term variants and therefore improve the recall performance in later NER stage. The dictionary used in the NER process is formed by a subset of UMLS and corresponding synonyms together with a pre-defined list of terms. The NER process uses noun phrases identified by POS tagging process and matches them to the dictionary. The NER also implements the NegEx (Chapman et al., 2001a) algorithm to recognise negation patterns.

cTAKES was evaluated with two different parameters: exact NER annotation and overlapping allowed NER annotation. With exact annotation, cTAKES achieved 80.1% in precision, 64.5% in recall and 71.5% in F-measure. With overlapped annotation allowed, it achieved 88.9% in precision, 76.7% in recall and 82.4% in F-measure. Although cTAKES has achieved reasonable performance, it can still be improved in many aspects:

- To implement disambiguation ability into NER process.
- To correctly resolve coordinated terms into multiple individual concepts.
- To recognise relationship between recognised concepts.

2.3.5 **NegEx**

NegEx is an algorithm developed to identify negation indications (Chapman et al., 2001a), which is widely integrated into many other clinical NLP tools such as MetaMap (Aronson and Lang, 2010) and cTAKES (Savova et al., 2010). In clinical narratives, mention of a clinical finding term or its descriptions do not always indicate the presence of the actual

finding. In fact, many of those frequently explicitly mentioned or described findings are usually absent in the patient, known as 'pertinent negatives' (Chapman et al., 2001b).

With low level of ambiguities in clinical sublanguages, NegEx assumes that occurrences of pertinent negatives in clinical text are limited to a limited number semantic types. And such occurrences have frequent and regular patterns. Therefore, NegEx is developed as a simple algorithm without sophisticated linguistic analysis. The algorithm uses two lexicon-based regular expressions to detect negations:

```
PATTERN1: negation phrase + [ 0-5 words] + UMLS term

PATTERN2: UMLS term + [ 0-5 words] + negation phrase
```

The above two regular expressions use two different lexicons respectively, 23 negation phrases for `PATTERN1` and 2 negation phrases for `PATTERN2`. Meanwhile, there is also a lexicon of 10 false negation triggers to remove ambiguous negations or double negatives, such as *not rule out*.

NegEx was evaluated on 1000 sentences containing 1235 occurrences of UMLS terms. This test set can be divided equally into two groups: group 1 contains NegEx negation phrases and group does not. Both groups are annotated by physicians in advance as gold standard. A much simpler algorithm that uses a very limited lexicon (*no*, *not*, *n't*, *denied*, *denies*, *without* and *ruled out*) is used as baseline method. The baseline method negates all UMLS terms between a negation phrase and the end of the sentence.

On group 1, NegEx achieved a precision of 84.49% and recall of 82.41%, compared with the baseline performance of 68.42% in precision and 88.27% in recall. Neither NegEx nor the baseline algorithm works on group 2 as there is no NegEx phrases contained in those sentences.

NegEx relies on the recognition of UMLS terms in text. In other words, NegEx does not work if no UMLS term is found in a sentence. Limitations in lexicons and regular expressions also affect its performance. NegEx cannot work on texts that do not contain negation phrases from pre-defined lexicons. The regular expression patterns limit only up to 5 intervening words can exist between the UMLS term and negation phrase, which has led to a decrease of its recall performance compared with the baseline method.

2.4 Summary

This chapter introduces some key concepts that are relevant to this research, including sublanguage and named entity recognition. We focused especially on their applications

in the clinical domain. Sublanguages have several characteristics that differentiate them from general languages. We also discussed substantial challenges to NLP applications in relevant sublanguage domains brought by these characteristics.

Some state-of-the-art clinical named entity recognition systems are reviewed, focusing on annual i2b2 shared task challenges. We selected a few top performance systems and briefly discussed their methods and performances.

With characteristics such as applicable semantic classification, controlled vocabulary, frequent co-occurrence patterns, contextual omissions, restricted abbreviations, knee MRI report narratives, which will be used for analysis in this project, is considered as a subpart of the clinical sublanguage. Therefore, it will also lead to corresponding difficulties.

Regarding challenges caused by sublanguage characteristics, we addressed through the following aspects:

- Semantic coverage of concepts in the reports was identified by applying MetaMap as an initial process that maps texts to UMLS, followed by classifications of these concepts by consulting with domain expert.
- Co-occurrence patterns and frequent term variants were identified using FlexTerm that groups similar phrases, and also through manual annotation.
- Meanwhile, manual annotation on part of the training set also helped reducing the disadvantage caused by the lack of annotated corpus.

With regard to NLP related problems, we addressed from the following aspects:

- Common NLP pre-processing steps, including tokenisation and segmentation, were performed using Stanford NLP
- NER was performed using a combination of dictionary-based and pattern-based approaches.
- Spelling mistakes were resolved by incorporating soft dictionary matching in dictionary-based NER using PathNER.

Patterns compiled in Mixup language were used to identify implicit concepts, as well as help resolving negated concepts and ambiguities.

Chapter 3 Clinical knowledge representation

In this chapter we provide a review of existing work related to formal knowledge representation, focusing specifically on ontologies as formal specification of domain knowledge commonly used to sharing and reuse such knowledge particularly in life sciences and clinical applications. We provide more details about a couple of knowledge resources of relevance to our own application, namely Unified Medical Language System (UMLS), Orchard Sports Injury Classification System (OSICS), Taxonomy for RehAbilitation of Knee conditions (TRAK) and Radiology Lexicon (RadLex). We also discussed some ontology related tools and applications to showcase use of ontology in the clinical domain.

3.1 Development of knowledge representation

To understand the real world, a human might need to go through the following steps to obtain enough information (Grenon, 2008):

- Observing around.
- Memorizing or making notes to help remember observations
- Analyzing observations and trying to structure them
- Experimenting with previous analysis results

The step of memorizing observations is a type of knowledge representation of a human. Human can solve a problem by using its own relevant information of the domain consciously or unconsciously. However, for a computer to achieve a solution for a problem, it needs to understand the problem first and then compute with relevant information. For example, if a doctor was told to treat a patient, he/she would use relevant medical knowledge to find out what injury the patient has and how serious it is. Then the doctor will be able to diagnose the patient and provide appropriate treatment. For a computer, if it were only told to treat the patient, it would not be able to do anything without proper representation of the exact problem and relevant knowledge background of the domain.

In order to enable a computer the ability of solving problems, it needs to go through a few more steps other than human actions as shown in ***Figure 3-1*** (Poole and Mackworth, 2010):

- Represent the problem using a language that can be reasoned by a computer
- Compute represented problem to generate relevant output

- Interpret generated output as a solution to the original problem

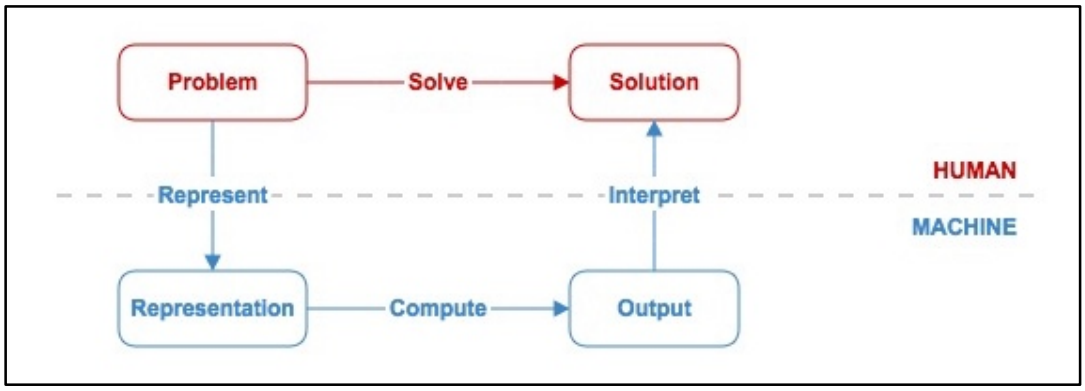


Figure 3-1 Problem solving steps (Poole and Mackworth, 2010)

Generally speaking, knowledge can be:

- Facts: attributes of an object or a group of objects
- Events: what, when, where, how
- Actions
- Meta-knowledge: rules of how to represent and apply knowledge

Knowledge must be represented in a form that computers can access and understand. A good representation scheme is a prerequisite. It should be (Poole and Mackworth, 2010):

- To cover every necessary detail aspect of the domain for problem solving
- To provide as accurate as possible representation of problems
- To represent with clear relationship between represented knowledge and original knowledge
- To be maintainable when there will be any change to the original knowledge
- To be efficient for computation

There are several general knowledge representation methods: logical representations, production rules and semantic networks.

Logical representation is a typical knowledge representation method and is the closest one to the format of natural language. Meanwhile, it has strict rules to allow communication with accuracy and no ambiguity. Its expressiveness also provides the ability to cover in-depth details. The higher the logic order is, the more expressive the logic representation is. However, it will also become more difficult to be reasoned (Gow, 2009).

Production rules are set of condition-action and premise-conclusion pairs. Production rules are also called if-then rules. An obvious advantage of production rules is that each rule is an independent item, so it can be added, modified or deleted accordingly. Any rule can be triggered at any time if its condition is met despite of other rules. As there may be multiple rules being triggered, a confliction resolution mechanism must be defined in advance (Lewis, 2010).

Semantic network is a type of graphical knowledge representations. It was first proposed by Quillian in 1969 as a hierarchical memory structure (Collins and Quillian, 1969). Semantic network is composed of nodes and connections, where nodes represent facts or concepts, and connections represent relations between nodes. The semantic network scheme has four fundamental components: lexical part, structural part, procedural part and semantic part. Lexical part decides what labels can be included in the representation to name nodes and connections. Structural part decides how those nodes are connected and what are those relations among them. Procedural part decides the access procedure to the semantic network. With definitions in this part, it will allow the creation of new nodes and connections, and the deletion of existed nodes and connections. End-users will also be able to derive answers to questions by following the procedure. Semantic part explains meanings underneath each node and connection (Bullinaria, 2005).

Using semantic networks as knowledge representation scheme has some obvious advantages:

- The ability to express the real world structure with its hierarchical structure and connections
- The semantics part allows accurate interpretations of real word concepts.
- Problems can be explicitly expressed. Therefore, they can be understood easily.

3.2 Ontology

The concept of ontology originated from the domain of philosophy dealing with existences. In the domain of information science, ontology was first mentioned by John McCarthy in the 1980s as a formalized general representation of a list of everything that exists (McCarthy, 1987). But it was Patrick Hayes who made the first ontology in his research of Naïve Physics (Hayes, 1983). In early 1990s, ontology became a standard component in knowledge systems (Neches et al., 1991). However, there is no standard definition of ontology.

Zúñiga (2001) proposed that a knowledge representation, which consists concepts for at least one specific domain and uses a specific formal language, can be considered as an ontology. Gruber defines an ontology as a group of formal concepts, relationships together with their definitions that forms knowledge of a domain (Liu and Özsu, 2009). Uschold et al. (1998) also provide a similar definition that the essentials of an ontology are a vocabulary of terms and relevant meaning. Concepts and relations in an ontology can be used as logical input and output for a knowledge system. Lightweight ontologies normally only have concepts and limited types of relationships. See ***Figure 3-2*** for an example. Constraints could also be added into an ontology to put restrictions on relationships and interpretations (Kalibatiene and Vasilecas, 2011).

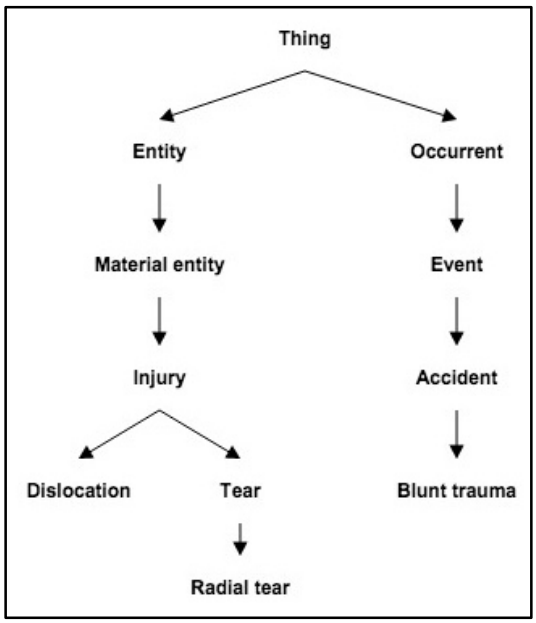


Figure 3-2 Part of a light-weighted ontology with only is_a type relationship

Ontology is also considered to be part of the Semantic Web, which is a computational meaningful form for Web content. Semantic Web is an extension of the existed Web, not a different version. It links relevant data together to enable machine understanding of linked information and relevance based search (Berners-Lee et al., 2001).

As the development of ontology, the term *ontology* is now widely used in many knowledge related approaches. It is now can be referred to any hierarchically presented structure in a computer language format (Grenon, 2008).

Ontologies can be constructed using many ontology design methods. However, as ontology can be count as some kind of software, it should also follows the IEEE Standard

1074-1997 guideline, which describes a general software development process (Fernandez-Lopez and Gomez-Perez, 2002; IEEE, 1998):

- **Project management activities**: define project initiation and expected output, plan for the whole design processes, and set up monitoring and control methods.
- **Pre-development activities:** study the domain where ontology will be used for, and review the possibility and feasibility applying the ontology.
- **Development activities:** identify system requirements and carry out actual design activities
- **Post-development activities:** apply the ontology with relevant knowledge system and identify possible future improvements
- **Integral activities:** evaluate the system, compile documentations and training manuals.

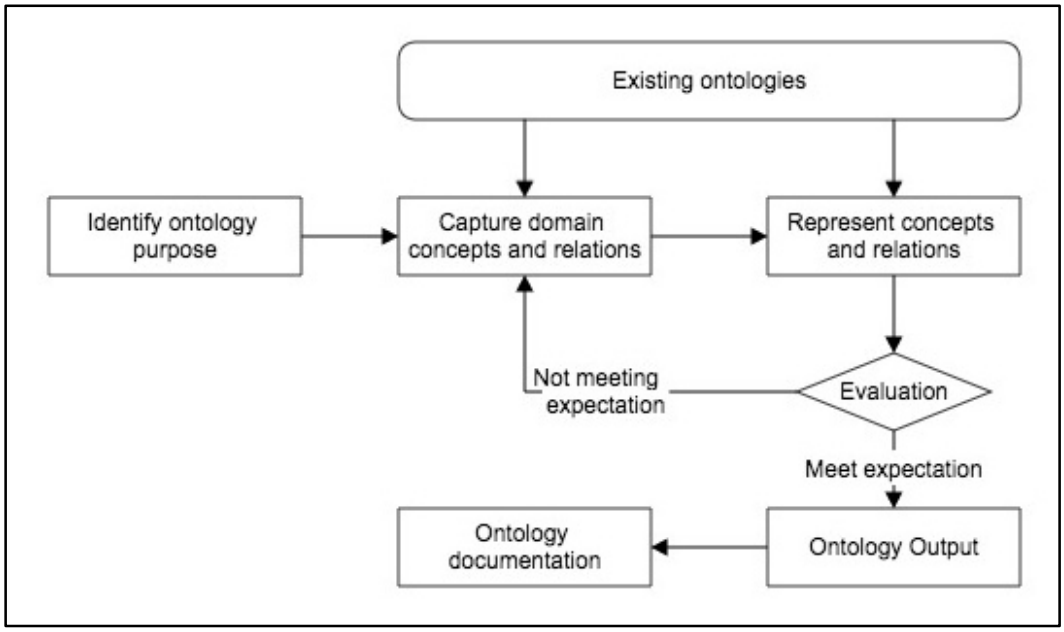


Figure 3-3 Skeletal ontology building method (Uschold et al., 1995)

There are many ontology building methods. Most of them follow the skeletal method (***Figure 3-3***), which is designed to provide a fundamental method that can be expanded into some future comprehensive ontology building methods (Uschold et al., 1995). The skeletal method initiates with identifying purposes and expectations of building the ontology. A clear understanding of its purposes and expectations would help to distinguish what concepts and relations need to be extracted from the domain and to be included in the ontology. Existing ontologies could be referenced to avoid duplicated extraction works and provide a guidance of categories of concepts and relations that need

to be extracted. It will also be able to help determining representation format. The ontology will be evaluated based on predefined purposes and expectations. If it does not meet the requirements, it will need to go back to look into extraction and representation again for better performance. Once satisfied the requirements, the ontology will be presented using selected representation format, provided with relevant documentations.

3.3 Available resources

There exist a lot of available clinical knowledge representations, such as MeSH (Lipscomb, 2000), OSICS (Rae and Orchard, 2007), RadLex, etc. Many of these clinical representations are included and indexed in the Unified Medical Language System (UMLS) (Bodenreider, 2004). Here we discussed some of these resources that were considered or referenced in this research.

3.3.1 **UMLS**

UMLS, short for *Unified Medical Language System*, includes one of the most widely referenced biomedical controlled vocabularies repository (Bodenreider, 2004). It also comes with a group of tools. MetaMap, as mentioned earlier, is one of these tools.

The repository is made up of three components: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The UMLS Metathesaurus is the key component that contains a large set of biomedical concepts. The Semantic Network includes 131 semantic types, which categorize all those concepts in the Metathesaurus (UMLS, n.d.). The SPECIALIST Lexicon includes common English vocabulary and biomedical vocabulary with relevant variants. It provides necessary lexical information for the NLP system (UMLS, n.d.).

The Metathesaurus is built from a large amount of health related thesauri, classification, code sets and various controlled vocabularies from different domains such as patient care, clinical research. With quarterly updates, it now covers over 150 vocabularies, such as NCBI taxonomy, Gene Ontology, MeSH, SNOMED CT and ICD-10-CM (UMLS, n.d.). For example, the NCBI taxonomy covers all names and classifications for all organisms from GenBank (Federhen, 2012). The SNOMED CT provides a set of standardized clinical representations (ihtsdo, 2014).

The UMLS was evaluated by Friedlin and Overhage on a collection of 3,000 chest x-ray reports to find out whether UMLS can be used to fully represent concepts derived from clinical narratives (Friedlin and Overhage, 2011). Among 5,975 unique noun phrases,

UMLS recognised 5,787 of them, which accounts 97% of total. Of all unique noun phrases, there were 1,687 (28%) of them being fully mapped to the Metathesaurus while 4,100 (69%) of them only being partially mapped. Apart from errors caused by MetaMap, the tool which was used to recognise and map terms, about 15% of all non-matched and partially matched phrases were because of missing relevant concepts in the Metathesaurus. The other major cause was that although some concepts were covered in the Metathesaurus, it was not represented properly to include necessary or commonly used synonyms.

3.3.2 **OSICS**

OSICS, short for *Orchard Sports Injury Classification System*, is a classification system developed for sports related injuries to support diagnosis and research. It was firstly developed in 1992 and the latest version is OSICS-10 (Rae and Orchard, 2007). OSICS-10 provides a vocabulary of sports injury related terms, definitions and basic hierarchical structure. According to definitions of ontology motioned earlier in ***Section 3.2***, OSICS-10 can also be considered as a simple ontology.

OSICS-10 is an improved version comparing to its predecessor OSICS-8. Version 8 was compiled using a 3-tier coding structure. In version 10, the coding structure was redesigned to become 4-tier to describe further detail for injuries, see ***Figure 3-4***. The first tier indicates the anatomical site and the second tier indicates pathology. The last two characters provide more detailed classifications based on information indicated from the first two characters. In this coding system, 'X' is used as non-specific code. 'X' has different meanings when it was used at different tier. It means location unspecified at tier 1 and injury unspecified at tier 2. For tier 3 and tier 4, it means no further classification can be identified.

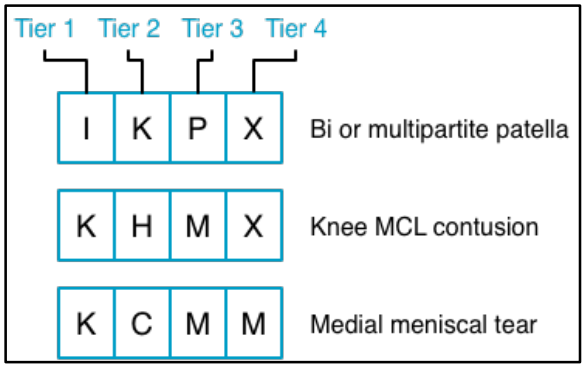


Figure 3-4 Example of the 4-tier coding structure used in OSICS-10 (Rae and Orchard, 2007)

For evaluation, Hammond et al. (2009) collected 335 diagnoses from professional sports unions and coded them using OSICS-10. All diagnoses were able to be coded using OSICS-10 coding structure, and in total there were 352 codes generated. The system uses a best-fit coding discipline, where it uses the highest tier of coding if possible. Although OSICS-10 managed to code all diagnoses, the coding system was not perfect enough. Among all coded diagnoses in tier 2, 49.2% of them could have be given a more specific tier 3 or tier 4 code instead due to the lack of specific classifications in the coding system.

3.3.3 **TRAK: Taxonomy for RehAbilitation of Knee conditions**

TRAK (K. Button et al., 2013), short for *Taxonomy for RehAbilitation of Knee conditions*, is an ontology developed for modelling information related with rehabilitation of knee conditions, as well as providing a framework for data collection that supports epidemiologic studies and clinical practices. This information includes classification of knee conditions, detailed knowledge about knee anatomy and an array of healthcare activities that can be used to diagnose and treat knee conditions (see ***Figure 3-5***). TRAK was developed using OBO ontology language following the design principles recommended by the Open Biomedical Ontologies (OBO) (Smith et al., 2007).

The original TRAK ontology consisted of 1,292 unique concepts and 518 relationship instances. Extra information provided with these concepts includes their preferred names, definitions, synonyms, unique identifiers and cross-references with external knowledge resources. TRAK ontology re-uses and cross-references with existing knowledge resources in its development process to keep consistency and avoid duplicated work. These external resources include UMLS Metathesaurus, Foundational Model of Anatomy and Orchard Sports Injury Classification System, etc.

Prior to the development of TRAK ontology, systematic literature review (K. Button et al., 2012) and questionnaire survey are integrated together for knee rehabilitation data collection. Collected terminologies are compared with UMLS for curation and standardisation, as well as obtaining extra information such as concept identifiers, synonyms, definitions and hierarchical structures. Such information, together with professional knowledge from domain experts, help with the definition of initial ontology structure and classification of rehabilitation concepts.

Figure 3-5 Upper-level hierarchy of TRAK with definitions of upper-level classes imported from the cross-referenced sources

The ontology also expands to include further information about knee conditions, diagnosis and treatment activities and anatomical locations for the purpose of supporting consistent and unambiguous data collection.

Knee conditions OSICS-10 is referenced and partially extracted to support systematic knee conditions classification. Each knee condition concept in OSICS-10 consists of two parts: the condition and affected anatomical location. These two-part concepts are separated and incorporated into the ontology under two semantic types, and also cross-referenced with UMLS for consistency. Most of the condition concepts are classified as *injury* or *disease*, which are classes re-used from the Ontology for General Medical Science (Robinson et al., 2008). Several additional categories from other existing resources (Grenon et al., 2004; Herre, 2010; Scheuermann et al., 2009) are also included for further subdivision of these concepts.

Diagnosis and treatment activities Concepts of diagnosis and treatment activities are obtained mostly by keyword-based searching from the UMLS. These keywords are proposed by domain experts who has professional knowledge. A few knee condition self-assessment questionnaires (Bellamy et al., 1988; Roos et al., 1998; Tegner and Lysholm, 1985) are also referenced to cover activities that cannot be found in the UMLS.

Anatomical locations Anatomical concepts are primarily from two sources: concepts derived from OSICS-10 and the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003). Concepts derived from OSICS-10 are those knee anatomical sites that affected by injuries and diseases. To achieve the generalisability, FMA is referenced and partially re-used. The structure of extracted FMA portion is also preserved.

With the vocabulary and taxonomy provide in TRAK, it can be used to support information retrieval system in the domain of knee injuries and rehabilitations. The largely cross-references terminologies and preserved general structure also enables it to be extensible to support other tasks in the domain. For example, the knowledge about knee anatomy, which is cross–referenced to a total of 205 concepts in the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003), is directly applicable to interpretation of reports describing knee MRI scans.

3.3.4 **RadLex**

RadLex (Langlotz, 2006) is a lexicon of radiology terminologies developed by the Radiological Society of North America (RSNA). RadLex is available from BioPortal (Whetzel et al., 2011). The original intentions of developing RadLex is to solve limitations coding in indexing and retrieving online image-based teaching materials, because of no available complete set of imaging terminologies from other resources at that time. These limitations are primarily caused by lack of details and human readable coding system. Therefore, RadLex is designed to provide a standard lexicon for all radiology-related information.

Although RadLex focuses on providing a standard lexicon, it has adopted a simple subsumption hierarchy, making it a basic ontology. There are 14 braches of different types of terms such as property, *radlex descriptor*, *anatomical entity* and *clinical finding*, etc. RadLex now contains34,446 active classes and a total of 58,065 terms.

The coverage of RadLex has been evaluated by various researchers. Hong et al. reviewed RadLex with 2,509 unique terms from 70 radiology reporting templates, and found an overall match rate of 67%, which consists 41% exact match and 26% partial match (Hong

et al., 2012). Woods and Eng (2013) also evaluated the coverage of RadLex in the domain of chest radiography reports. Among 339 unique terms, RadLex achieved an overall match rate of 62%. Woods and Eng's research also revealed a significant shortcoming of RadLex on lack of inclusion of many frequently mentioned medical procedures.

3.4 Use of ontologies

Ontologies are developed to model, share and reuse knowledge within a specific domain. Many ontologies, relevant tools and ontology-based applications have been developed for various domains, including clinical, linguistic, chemical, etc. Stevens, Goble and Bechhofer classified usages of ontologies into three types, which are domain-oriented, task-oriented and generic uses. These usages are not completely isolated from each other. Usually, ontologies and ontology-based applications would have mixed usages of these three types.

3.4.1 **Ontology repositories**

There exist two major online ontology repositories: the NCBO BioPortal (Whetzel et al., 2011) and the EBI Ontology Lookup Service (OLS) (Côté et al., 2006).

NCBO BioPortal

NCBO BioPortal is a web service that provides online access to various biomedical ontologies represented using popular ontology languages such as OWL and OBO. After years of improvement, BioPortal now contains 446 ontologies from various domains, including health, human and anatomy as the top 3 sources (NCBO, 2013). It provides users convenient accesses to ontologies for term information retrieval and the capability to be implemented into other software applications through its public APIs (Whetzel et al., 2011). The BioPortal web services and its APIs are both developed based on RESTful Web services. Information can be retrieved includes the ontology metadata, all terms from an ontology, detail information of each term and cross mapping among all available ontologies on BioPortal.

Ontology Lookup Service

The OLS was one of the earliest public online ontology browsers. It now contains 148 ontologies (EMBL-EBI, 2016). There are overlapped coverage of ontologies compared with BioPortal. However, some ontologies are only available at one of the two repositories. OLS also provides functions as ontology querying and visualisation, as well as a RESTful API based that allows retrieval of meta-data from ontologies and terms.

3.4.2 **Ontology tools**

There are many tools available for developing, editing and querying ontologies, such as Protégé (Stanford Centre for Biomedical Informatics Research, n.d.), OBO-Edit (Day-Richter et al., 2007) and OntoCAT (Adamusiak et al., 2011).

Protégé

Protégé is an open-source general purpose ontology editor developed at the Stanford University. It also provides a framework that can be used to build ontology-based applications. Benefit from the implementation of the OWL API (Horridge and Bechhofer, 2011), Protégé supports a wide range of ontology languages such as OWL, OWL/XML, RDF/XML and OBO Flat file format.

OBO-Edit

Compared to Protégé as a general purpose ontology editor, OBO-Edit is designed specifically for biology related uses. It has limited number of supported formats with optimised support for its OBO format. With features of human-readable and more concise structure compared with XML, the OBO format is recommended by the Gene Ontology Consortium (The Gene Ontology Consortium, 2015).

OntoCAT

OntoCAT (Adamusiak et al., 2011) is a software package that provides ontology search using various programming languages such as Java, R and RESTful API. It is developed for access and search ontologies in OWL and OBO format from various locations including BioPortal, OLS and local storages. OntoCAT allows searching for terms, synonyms, definitions and annotations within an ontology. It also allows hierarchy based retrieval to obtain child and parent terms, relationships and paths towards the root term.

3.4.3 **Ontology-driven applications**

Ontologies are often integrated for tasks in the clinical domain, such as concept extraction, data integration, document classification and clinical decision support system, etc. Here we selected only a few recent ontology-based applications only to showcase some typical uses.

Rahimi et al. (2014) developed a Diabetes Mellitus Ontology (DMO) and a corresponding ontology-based algorithm to diagnose and manage patients with diabetes. The DMO ontology provides an in-depth well covered knowledge base to represent clinical data

extracted from electronic health records, and provide support for making clinical decisions about diabetes care, as well as support for researches about diabetes. The algorithm is based on ontology queries using Semantic Protocol and RDF Query Language (SPARQL). Given achieved sensitivity of 100% and specificity of 99.88% in identify reason for visit, 96.55% and 98.97% in medication, 15.6% and 98.92% in pathology, the algorithm has been proved as sufficient to support clinical decisions. Manzoor et al. (2015) also built a clinical decision support system based on an automatic constructed ontology to predict high-risk pregnant woman.

Mate et al. (2015) design an ontology-based approach for data integration between clinical and research. With this approach, it addresses difficulties in researchable reuse terms from electronic medical records (EMR) that are not code and linked to terminologies. Three ontologies are created for the whole data integration process: clinical source ontology, mapping process ontology and research target ontology. Unlike other approaches, an ontology is used to store semantic mapping rules and later to be automatically converted into SQL queries to facilitate the processes of database-ontology conversion.

Osborne et al. (2009) utilise the Disease Ontology (Schriml et al., 2012) for human genome annotations. The annotation is performed on extracted gene-disease relationships from the GeneRIF (Jimeno-Yepes et al., 2013) database. The annotation result is evaluated against the Homayouni gene collection (Homayouni et al., 2005) and compared with result from the Online Mendelian Inheritance in Man (OMIM) (Rashbass, 1995), which is an online gene catalog. The ontology-based approach achieved a precision of 97% and recall of 91% compared with 98% in precision and 22% in recall for OMIM.

3.5 Summary

This chapter introduces the concept of knowledge representation and its development. We also introduced ontology, which is a sub-concept of knowledge representation. Among all available clinical knowledge representations, we selected and discussed a few that are integrated or referenced in this research, focusing on UMLS, OSICS, TRAK and RadLex.

We also reviewed some contemporary ontology-related tools and applications to showcase how ontologies can be stored, edited and integrated.

This thesis is involved with many clinical knowledge representations, including several ontologies. The TRAK ontology was particularly studied and expanded in this project.

For the expansion of the TRAK ontology, OSICS, RadLex and UMLS are primary knowledge sources that we referenced for the inclusion of possible terms and hierarchical structure.

These clinical knowledge representations were used on the following aspects:

- UMLS and OSICS were used as dictionaries to recognise terms from knee MRI report, so that we could identify which terms are not included in the original TRAK ontology. Semantic relationships also obtained.
- We also manually extracted terminologies from RadLex and MEDCIN (a UMLS subset) to extend the coverage of possible inclusions from our dictionary-based approach. The hierarchical structure of RadLex descriptor branch was also incorporated into the expanded TRAK as complementary inclusions on descriptors in addition to UMLS.
- The expanded TRAK ontology itself also acted as a dictionary to drive the information extraction process. It also provided semantic relationships for ambiguity resolutions.

Chapter 4 Annotation and analysis of MRI reports

In this chapter we describe a corpus of text documents used for training and testing. We describe the type of documents considered, the provenance of data and basic statistical properties of the corpus. To illustrate its semantic coverage, we report the distribution of concepts and semantic types mentioned in the corpus. These mentions were extracted automatically from the corpus using the Unified Medical Language System as a reference point. We proceed by describing the linguistic pre-processing performed on the corpus. We finish the chapter by describing manual annotation of the corpus as part of preparing it for training and testing purposes.

4.1 Data source

In this section we motivate the use of MRI reports in our system. Magnetic resonance imaging (MRI) is a technique that uses a strong magnetic field to visualise organs and structures of internal body by detecting radio wave energy pulses emitted by tissues (WebMD, 2014). MRI may reveal problems that cannot be seen with other imaging methods such as X-ray, ultrasound or CT. For example, compared with X-ray, MRI provides better results on soft tissue visualisation. Therefore, MRI is sometimes used to provide more details of a suspect problem seen on other imaging results. Meanwhile, MRI could also produce three dimensional images so that tissues can be viewed from different angles (Kwong and Yucel, 2003).

For knee pathology diagnosis, together with patient medical history and other examinations, MRI could become an effective assistant tool that provides increased accuracy in diagnosis and guiding treatments (Grover, 2012; Konan et al., 2009; Pompan, 2012; Wenham et al., 2014; Yan et al., 2011). For example, meniscus tear is a common type of knee injury with 22.4% prevalence among all soft tissue injuries for patients attending a trauma department (Clayton and Court-Brown, 2008). With assistance of MRI, the accuracy of individual test based meniscus diagnosis can be increased to 96% from 74% previously without MRI (Konan et al., 2009). In a previous study of meniscus tear diagnosis (Yan et al., 2011), comparison result of improved diagnostic performances in ***Table 4-1*** shows that combining MRI with patient medical history and other examinations brings higher accuracy, sensitivity and negative predictive value among all three independent clinical diagnostic factors (giving away, locking and McMurray's test) for meniscus tear diagnosis, ***see Table 4-1***. With this study, it proves that MRI should be considered as a recommended diagnosis procedure for knee injuries.

Table 4-1 Comparison of diagnostic values for meniscus tear with and without MRI (Yan et al., 2011)

Diagnosis factors	Accuracy	Sensitivity	Specificity	PPV*	NPV*
Giving way+	49.2%	43.5%	84%	94.4%	19.4%
	88.3% (MRI)	95.7% (MRI)	74.2% (MRI)	87.5% (MRI)	90.2% (MRI)
Locking+	60.9%	55.2%	96%	98.8%	25.8%
	89.9 (MRI)	97.4 (MRI)	75.8% (MRI)	88.4% (MRI)	94% (MRI)
McMurray's test+	76%	75.8%	76.9%	95.1%	35.1%
	89.4% (MRI)	97.4 (MRI)	74.2% (MRI)	87.7% (MRI)	93.9% (MRI)

* PPV = Positive predictive value, NPV = Negative predictive value
+ Giving way, lock and McMurray's test are three key factors that helps to diagnose meniscal tear

Another recent research has also highlighted the importance of MRI in symptomatic early knee osteoarthritis diagnosis and treatment planning. MRI will be recommended to the patient if it has some symptoms from clinical examinations but has a normal X-ray result. The MRI report will then be used as an evidence to determine whether the patient needs surgical or nonsurgical treatment. MRI could also greatly decrease the need for costly and invasive arthroscopy diagnosis (Luyten et al., 2012; Wenham et al., 2014).

Clinical radiology images, including MRI, are usually accompanied with corresponding reports, which provide professional interpretations of images. These reports also relate patient symptoms and images together to provide suggested diagnosis (Royal College of Radiologists, 2006).

For research purposes, MRI images and reports are also important supporting evidences for knee pathology epidemiologic studies (Guermazi et al., 2012; Roemer et al., 2011). Particularly in the longitudinal studies of knee osteoarthritis, MRI has proved its importance that it could identify early lesions, which are not shown on other radiographic reports and are believed to be indicative of subsequent osteoarthritis development (Javaid et al., 2010; Pessis et al., 2003).

However, in recent studies, it is quite usual to have false findings due to low statistical power, bias, number of studies on the same question and the ratio of true to no relationships among probed relationships (Ioannidis, 2005). One major cause underneath these concerns is the low sample size of studies, although the relationship is not simple or proportional. Normally large scale studies also require more funding and personnel resources (K. S. Button et al., 2013). Therefore, it is not surprising to have size limited (hundreds normally or even dozens in rare cases) epidemiologic studies due to the complexity and cost of manual interpretation of medical notes. Large datasets could be analysed within a short time using computers. Therefore, if the interpretation process of

MRI reports could be automated, it would be able to eliminate the obstacle of sample size limitation in relevant studies those require manual annotation of complex data such as images.

In previous studies, it has been proved that clinical narratives such as pathology and radiology reports could provide valuable diagnostic information (Mohanty et al., 2007). It has also been proved that NLP approaches could be applied to those narratives to extract structured information in a recent cancer-related research (Spasić et al., 2014). This means that it is in principle feasible to automate interpretation of clinical narratives such as those found in imaging reports.

4.2 Data provenance

A dataset, which consists of 1,468 MRI scan reports, was obtained from the Cardiff and Vale University Health Board (C&V UHB). During a period of 11 years and 5 months from January 2001 to May 2012, there were 6,382 individual patients visited the Acute Knee Screening Service at the Emergency Unit of the Cardiff and Vale University Health Board (C&V UHB). 1,657 of these patients were referred to take an MRI scan. A radiology report was provided by a radiologist from a group of five following each MRI scan. Each report includes a professional interpretation of its MRI scan image result together with previous clinical assessment if applicable. These reports were stored securely in a C&V UHB managed database, which was originally designed for service evaluation and auditing purposes. Not all of those 1,657 patients who were referred to take an MRI scan have actually attended. Therefore, only a total of 1,468 MRI reports were identified and extracted from the database. These 1,468 records formed the dataset used in this study. All these reports were made anonymised prior to their release, i.e. any information that can identify a patient or radiologist was removed from these reports. The dataset was transferred using an encrypted memory stick. All devices that contain these records were stored securely in locked rooms. To carry out research using these reports, we have obtained ethical approval from the South East Wales Research Ethics Committee, reference number 10/MRE09/29.

The Radiological Society of North America (RSNA) has created a library of clear and consistent report templates to improve reporting practices and help radiologist to provide high-quality radiology reports more efficiently. It also provides a template for knee MRI reports, including the following sections: procedure, clinical information, comparison, findings and impression (Radiological Society of North America, 2012). However,

structures from reports that we have obtained do not match this template. See ***Textbox 4-1*** for a sample report.

```
HISTORY Twisting injury, ACL rupture and medial meniscal tear
MRI RT KNEE There is some oedema within the ACL but the fibres are intact
and this would represent sprain or partial tear. The PCL is buckled but
it is intact. Both menisci are intact. Normal collateral ligaments. There
is a small amount of soft tissue oedema just posterior to the proximal MCL
insertion. It is probalbly due to direct trauma. There is no evidence of
a meniscal cyst. There is a small popliteal cyst. The popliteus tendon and
posterolateral corner are intact. Normal articular cartilage and extensor
tendons.
CONCLUSION ACL sprain/partial tear. Both menisci are intact. Soft tissue
oedema from direct trauma posterior to the proximal MCL insertion. The
collateral ligaments are intact.
```

Textbox 4-1 An example of MRI report structure

A typical MRI report from our dataset consists of three sections: history, latest scan result and conclusion. These sections are usually indicated using upper case phrases, such as *HISTORY*, *MRI RIGHT/LEFT KNEE*, *INDICATION*, *FINDINGS* and *CONCLUSION*. The history section describes previous diagnosis record. The latest scan result section usually starts with *MRI RT/RIGHT KNEE* or *MRI LT/LEFT KNEE*, containing the actual description of the corresponding knee that was scanned. The *CONCLUSION* section provides summarized description based on previous diagnosis record and latest scan result. History and conclusion parts are not necessarily included in every report.

4.3 Statistical properties

The dataset consists of 1,468 individual MRI reports. The size of the overall dataset is 1,002KB. The dataset has 13,991 sentences, of which in total 178,931 tokens. Among those tokens, there are 3,277 distinct tokens and 2,681 distinct stems. An average size of an individual MRI report is 0.68KB(±0.40KB) with an average of 9.53(±5.13) sentences and 110.81(±64.60) tokens. The overall dataset was separated into a training set and a test set. The test set was annotated manually by a domain expert to create a gold standard. As this task required specialised expertise and considerable manual effort, which was not readily available, we randomly selected only 100 MRI reports from the whole dataset. Documents selected for the test set were then removed from the dataset and were not considered while building the system. The test set was to be used for system evaluation at later stage so it was kept unseen. After removing the 100 documents selected to become

the test set, the remaining 1,368 reports formed the training set for building the system. ***Table 4-2*** shows some statistics of the training set.

Table 4-2 Training set statistics

	Overall	Average	Standard Deviation
Sentences	13,051	9.54	5.09
Tokens	166,824	121,948	69.72
Distinct Tokens	3,188	-	-
Distinct Stems	2,611	-	-
Size (KB)	934	0.68	0.41

As there is no existing annotated corpus in this domain, to achieve higher efficiency in system development, a subset of 100 documents was also randomly selected from the training set to form an active development set. This development set will be intensively used at later stage for manual annotation and performance evaluation during the development process.

4.4 Semantic coverage

To figure out what information was included in these reports, we started with term identification. We considered both UMLS and RadLex as external resources for this process. However, the result from a previous research in which Wang and Vall (2011) compared coverages of RadLex and UMLS on radiology reports, shows that UMLS has much better coverage than RadLex on radiology reports. They were both applied on a set of de-identified radiology reports that contain 176,091 noun phrases. Only 2.32% of phrases can be exactly mapped to RadLex, and only 50.53% can be partially mapped. On the contrary, 10.40% of phrases can be exactly mapped to UMLS and 85.95% can be partially mapped. Meanwhile, the vast majority (68%) of existing TRAK concepts were originally cross–referenced to the UMLS in an attempt to standardise the TRAK terminology and facilitate its integration with other terminological sources (Button et al. 2013). During the initial development of TRAK, the UMLS was searched collaboratively by a physiotherapist (who was both practitioner and researcher) and an informatician to obtain concept identifiers, synonyms and definitions, where such information was available. Therefore, we decided to use UMLS instead of RadLex to search for potential terms to be included into TRAK for its better coverage on radiology reports and to keep consistency with the original development of TRAK. MetaMap is a software tool for recognizing mentions of the UMLS concepts in text (Aronson and Lang, 2010). We applied MetaMap to the training set that consists of 1,368 MRI reports, and obtained

Concept Unique Identifiers, preferred concept names and semantic types from the candidate output. See ***Textbox 4-2*** for a sample output.

```
Candidate:
  Score: -861
  Concept Id: C0022742
  Concept Name: Knee
  Preferred Name: Knee
  Matched Words: [knee]
  Semantic Types: [bpoc]
  MatchMap: [[[2, 2], [1, 1], 0]]
  MatchMap alt. repr.: [concept start: 2, concept end: 2]
  is Head?: true
  is Overmatch?: false
  Sources: [FMA, HL7V2.5, LCH, UWDA, AOD, CHV, CSP, SNMI, SNOMEDCT]
  Positional Info: [(9, 4)]
  Pruning Status: 0
```

Textbox 4-2 An example of MetaMap candidates output

rowid	source	text	matched_...	candidate...	concept_id	concept_...	preferred...	semantic_...	head	overmatch	position
1	1	Knee	knee	-861	C0022742	Knee	Knee	bpoc	true	false	[(9,4)]
2	1	Knee	knee	-861	C0022745	Knee	Knee joint	bsoj	true	false	[(9,4)]
3	1	Knee	knee	-861	C1283838	Knee	Entire kne...	blor	true	false	[(9,4)]
4	1	Knee	knee	-861	C1963703	Knee	Knee regi...	blor	true	false	[(9,4)]
5	1	HISTORY	history	-694	C0019664	History	Recording...	ocdi	false	false	[(0,7)]
6	1	HISTORY	history	-694	C0019665	history	Historical ...	inpr	false	false	[(0,7)]
7	1	HISTORY	history	-694	C0262512	History,NOS	History of...	orga	false	false	[(0,7)]
8	1	HISTORY	history	-694	C0262926	History	Medical H...	fndg	false	false	[(0,7)]
9	1	HISTORY	history	-694	C1705255	History	Concept ...	cnce	false	false	[(0,7)]
10	1	HISTORY	history	-694	C2004062	History	History of...	fndg	false	false	[(0,7)]
11	1	HISTORY	historical	-623	C1552658	Historical	Historical ...	idcn	false	false	[(0,7)]
12	1	HISTORY	historical	-623	C1552723	Historical	Historical ...	idcn	false	false	[(0,7)]
13	1	giving	given	-966	C1442162	GIVEN	GIVEN	cnce	true	false	[(14,6)]
14	1	giving	given	-966	C1550718	given	Entity Na...	idcn	true	false	[(14,6)]
15	1	giving	give	-966	C1947971	Give	Give - do...	ftcn	true	false	[(14,6)]
16	1	giving	given	-966	C3244317	given	Given name	inpr	true	false	[(14,6)]
17	1	MRI KNEE	mri,knee	-829	C0412714	MRIknee	Magnetic ...	diap	true	false	[(39,3),(4...
18	1	KNEE	knee	-795	C0022742	Knee	Knee	bpoc	true	false	[(46,4)]
19	1	KNEE	knee	-795	C0022745	Knee	Knee joint	bsoj	true	false	[(46,4)]
20	1	KNEE	knee	-795	C1283838	Knee	Entire kne...	blor	true	false	[(46,4)]

<< < 1 to 100 of 239839 > >>

Figure 4-1 MetaMap output saved in database

Semantic types from the MetaMap output provide a consistent categorization of all UMLS concepts. Among all 239,839 candidates suggested in MetaMap output shown in ***Figure 4-1***, there were a total of 121 different semantic types and combinations. Using Pareto principle (i.e. 80:20 rule) as a guideline, we assumed that most observations made by radiologist fell into top 20% frequently occurred semantic types (Clauset et al., 2009), see ***Figure 4-2***. ***Table 4-3*** lists out top 20% frequently occurred semantic types. Furthermore, these top 20% semantic types could be classified into 6 categories: medical condition, condition extent, body part, tissue, functional mechanism, and spatial qualifier (***Table 4-4***).

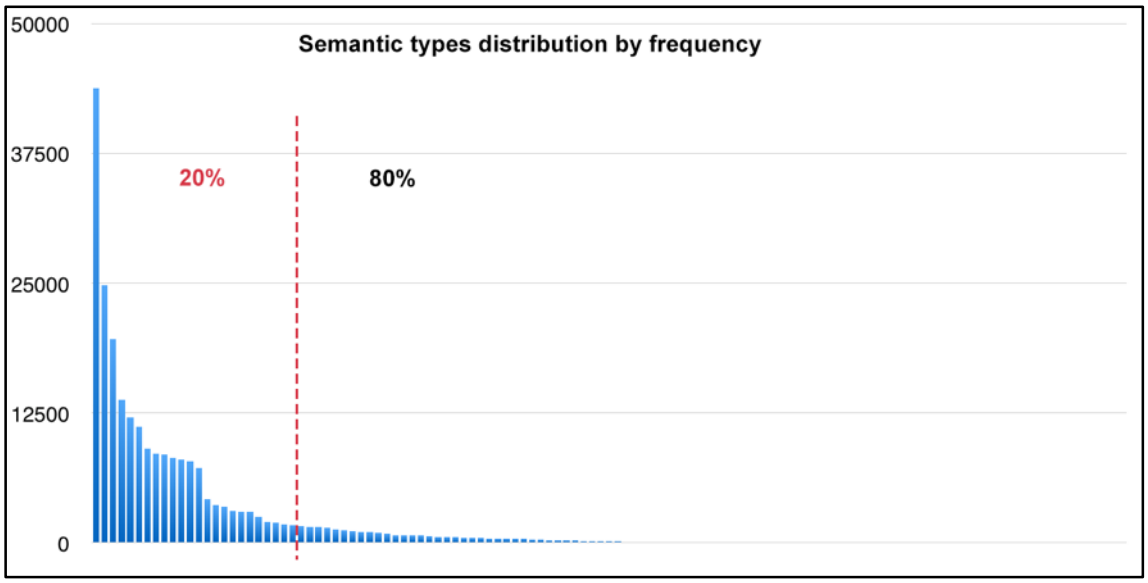


Figure 4-2 Semantic types distribution by frequency

Table 4-3 Top 20% frequently occurred semantic types

Semantic Types	Frequency
Body Part, Organ, or Organ Component	43795
Spatial Concept	24843
Qualitative Concept	19659
Quantitative Concept	13808
Injury or Poisoning	12111
Functional Concept	11191
Intellectual Product	9116
Body Location or Region	8625
Tissue	8535
Idea or Concept	8216
Body Substance	8033
Finding	7855
Body Space or Junction	7181
Medical Device	4209
Diagnostic Procedure	3648
Disease or Syndrome	3482
Bacterium	3098
Body System	3003
Pathologic Function	2995
Eukaryote	2504
Gene or Genome	2040
Clinical Attribute	1924
Organism Attribute	1748
Laboratory Procedure	1689

Based on the semantic types classification above, we defined an information extraction template that consists of the following slots: finding, finding qualifier, negation, certainty, anatomy, and anatomy qualifier. Medical condition corresponds to what we call finding slot. Condition extent was split into finding qualifier, negation and certainty. Body part

corresponds to anatomy slot. Tissue, functional qualifier and spatial qualifier were merged into anatomy qualifier.

Table 4-4 Classification of semantic types

Classification	Definition	Example
Medical condition	Injury, disease or symptom.	loss, tear, bruising, oedema
Condition extent	The extent of a medical condition.	large, early, possible, not found
Body part	An anatomical entity (body part) affected by the medical condition.	femoral condyle
Tissue	Specific tissue of the body part affected by the medical condition.	cartilage, bone marrow
Functional qualifier	Further specifies the body part in terms of its functionality.	weight bearing surface, extensor mechanism
Spatial qualifier	Further specifies the body part of the medical condition.	posterior horn, centre, distal third

4.5 Annotation

For the purpose of system development and evaluation, the test set and a limited subset of the training set were annotated manually with tags corresponding to slots and relationships from the predefined template.

The test set was annotated and kept unseen from the system designer to ensure that the system was designed unbiased. The annotated test set was used to create a golden standard for system evaluation. The annotated training subset of 100 documents was used to test the system during development and meanwhile to provide manually annotated terms for ontology expansion.

4.5.1 **Tag set**

Based on the distribution of top 20% of semantic types in the corpus (see ***Table 4-3***), we have chosen a set of six tags, which are described in ***Table 4-4***. In essence, we grouped related semantic types into a general category. For example, concepts from semantic types such as Finding, Injury or Poisoning, Disease or Syndrome, Pathologic Function, etc. are covered by the *finding* tag. Similarly, Body Part, Organ, or Organ Component, Body Location or Region, Body Space or Junction, etc. represent anatomical sites and as such are simply annotated using the *anatomy* tag.

We introduced a *certainty* tag to annotate the confidence with which the radiologist diagnosed a given *finding*. The *negation* tag is used to explicitly annotate negative findings. We also introduced two relationship tags: *applies_to* and *observed_in*.

observed_in connects *finding* to *anatomy*. And *applies_to* is used on the following relation pairs:

- *finding_qualifier - finding*
- *anatomy_qualifier - anatomy*
- *certainty - finding*
- *negation - finding*

Table 4-5 Semantic type classifications to annotation tags conversion interpretation and examples

Tag	Interpretation	Example
finding	Clinical manifestations observed by a radiologist.	intact, lesion, cyst, tear
finding qualifier	Property of a finding.	partial, complex, thinning
anatomy	Anatomical entity affected by the given finding.	lateral meniscus, ACL, bone marrow
anatomy qualifier	Further anatomical localisation.	posterior, inferior, posterolateral
certainty	Certainty with which a radiologist diagnosed the given finding.	seen, appear, evidence
negation	Indication of a negative finding.	no, not, without

4.5.2 **Guidelines**

Given the above tag set, the annotation process was carried out under the following guidelines:

- A phrase that describes a clinical finding (e.g. injury, disease or symptom) will be tagged as *finding*.
- If there is any phrase that further specifies a *finding*, it will be tagged as its *finding_qualifier*, and will be connected using tag *applies_to*.
- A *finding* tag can be linked with multiple *finding_qualifier* tags.
- An anatomical site directly affected by a given finding will be tagged as *anatomy*, and the *finding* will be connected to the *anatomy* using tag *observed_in*.
- A *finding* tag can be linked with multiple *anatomy* tags.
- If there is any further specification that describes the *anatomy*, it will be tagged as *anatomy_qualifier* and will be connected using tag *applies_to*.
- An *anatomy* tag can be linked with multiple *anatomy_qualifier* tags.
- Any phrase that describes the radiologist's certainty in diagnosing the given finding will be tagged as *certainty*.
- A *finding* tag can be linked with multiple *certainty* tags.
- Any phrase that indicates that a given *finding* is negative will be tagged as *negation*. The absence of a negation tag indicates that the *finding* is positive.

- A *finding* tag can be linked with at most one *negation* tag.

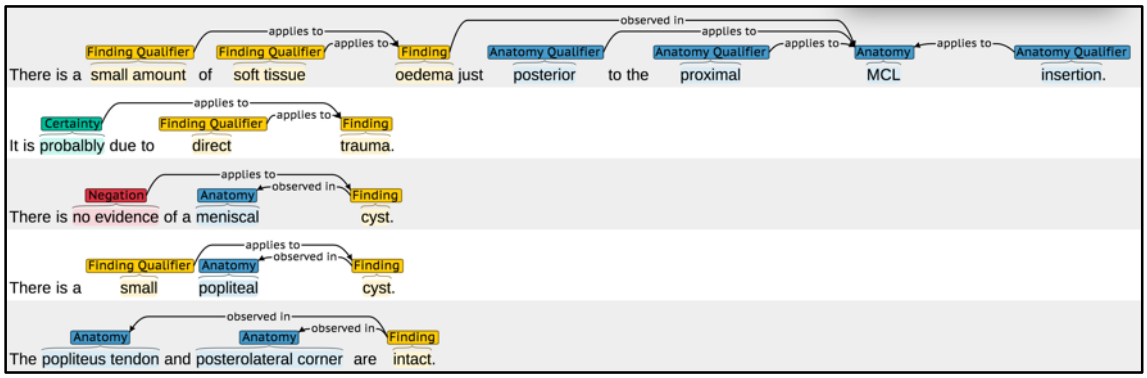


Figure 4-3 An example of annotated report featuring the use of different tags and relationships

Figure 4-3 provides an example of annotated report. Manual annotation was performed using BRAT, a web-based annotation tool (Stenetorp et al., 2012). It provides a highly intuitive graphical annotation interface. New annotation can be added by double-clicking on a word or drag-selecting a phrase. Meanwhile, relationship instances can be added by dragging one annotation onto another. The back end is also highly configurable with functions such as customized visualizing scheme, external search.

4.5.3 **MetaMap performance**

We also evaluated the performance of MetaMap. We applied MetaMap to the development set and compared its output with corresponding annotations from the two annotators respectively, see ***Table 4-6***. The performance evaluation is based on exact match, where partial match in a phrase does not count as correct extraction. The precision and recall scores of MetaMap performances on development set indicates more than 30% of recognised terms are incorrect and less than 60% terms have been recognised.

Table 4-6 MetaMap performances on development set (Exact match)

	Precision	**Recall**
Annotator A	0.665	0.562
Annotator B	0.69	0.565

4.5.4 **Gold standard**

The main purpose of creating a gold standard was to test the performance of the system on unseen data. In order to create a gold standard, the test dataset was annotated manually by two independent annotators, i.e. the student and supervisor, based on their expertise gained during the development of the TRAK ontology.

Fleiss' Kappa coefficient (Fleiss, 1971) was used to measure the inter-annotator agreement to show the reliability of the annotation work (Artstein and Poesio, 2008). Fleiss' Kappa can be calculated using the following equation:

$$\mathcal{K} = \frac{A_o - A_e}{1 - A_e}$$

A_o represents observed agreement and A_e represents expected chance agreement. The observed agreement is the percentage of annotations that both annotators agree. The expected chance agreement is calculated on the assumption that random assignment of categories to items, by anyone of the two independent annotators, is governed by the distribution of items among categories in the actual world (Artstein and Poesio, 2008).

The Fleiss' Kappa for annotation on the development set was calculated using an online tool (Geertzen, 2012). The calculated results were the observed agreement $A_o = 0.87$ and the expected chance agreement $A_e = 0.26$. Therefore, Fleiss' Kappa coefficient was calculated to be $\mathcal{K} = 0.825$. Following the guidance of Fleiss' Kappa value interpretation (Landis and Koch, 1977), $\mathcal{K} = 0.825$ indicates almost perfect agreement, see ***Table 4-7***. ***Figure 4-4*** also provides a marginal percentage distribution of tags used in annotation, which indicates the frequency of using each tag for each annotator. Such inter-annotator agreement shows that it was a reliable annotation.

Table 4-7 Fleiss' Kappa coefficient value interpretation (Landis and Koch, 1977)

$\mathcal{K}$	Interpretation
< 0	Poor agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

A gold standard was created by the third annotator, a clinician with a specialism in treating knee conditions, who independently resolved the inter–annotator disagreements, ensured the consistency of annotations and mapped individual annotations of text spans to the corresponding concepts in the TRAK ontology. Gold standard annotations were converted to filled IE templates represented as JSON objects in order to support their comparison to system output during evaluation. ***Figure 4-4*** shows the distribution percentage of annotation tags used in the development set and the gold standard. Distributions of tags used by the two annotators in the development set are similar with tag distribution in the gold standard. The tag *N/A* represents missing annotation or a term

that cannot be tagged using these predefined tags. Compared with annotations on the development set from the two annotators, the gold standard is annotated by a clinical professional. Thus, it has relatively lower level of *N/A* tags.

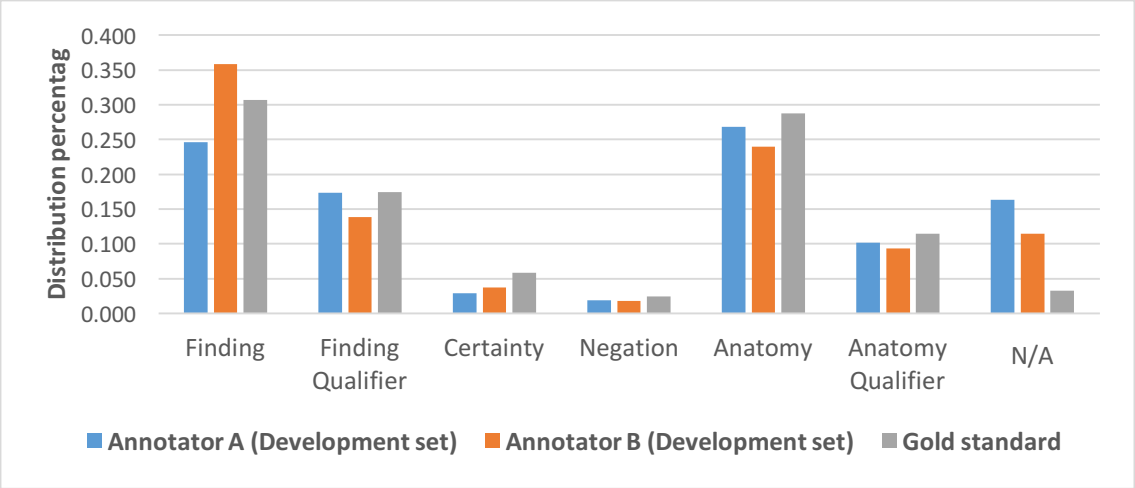


Figure 4-4 Distribution percentage of annotation tags in development set and test set

Chapter 5 Rapid ontology development strategies

In this chapter we describe an existing ontology that used as a formal representation of domain-specific knowledge in order to drive the information extraction process and map information extracted from textual space to conceptual space. We proceed with the description of processes used to make the ontology fit for this purpose. We present four strategies used to efficiently and systematically expand the ontology's domain coverage and its vocabulary. To achieve this, we devised four strategies to extract concepts and terms from available data and knowledge sources.

5.1 Strategies

In order to support semantic interpretations of clinical narratives found in MRI reports, we need to expand the original TRAK ontology to include all relevant MRI specific concepts, such as *hyaline cartilage abnormality*, *bone bruise*, *cyclops lesion*, etc. In order to support NLP applications of the ontology, its vocabulary also needed to be expanded to include term variants commonly used in MRI reports. Some term variants are confined to a specific clinical sublanguage (Hripcsak et al., 2002) and as such are typically underrepresented in standardised medical dictionaries such as those included in the Unified Medical Language System (UMLS) (Bodenreider, 2004). For example, *collateral ligament* was found to have no other synonyms in the UMLS. Yet, *collateral ligaments* are colloquially referred to as *collaterals* in clinical narratives. Thus, out of 37 references to collateral ligaments in the training dataset, six (i.e. 16%) accounted for this informal variant of the term.

Four strategies have been devised for systematic expansion of the coverage of the TRAK ontology, of which three are data-driven. The purpose of using data-driven strategies is to ensure that the ontology is appropriate for intended NLP applications on such data. Each data-driven strategy utilizes a different approach to extract relevant terminologies from the dataset either manually or automatically. To avoid over-fitting the ontology caused by limited data source in data-driven strategies, and to provide an initial taxonomic structure to incorporate new concepts, the fourth strategy integrates terminologies and references structures from relevant knowledge bases. These four strategies are:

1) Dictionary-based term recognition
2) Automatic term recognition
3) Manual data annotation

4) Manual dictionary search

The four strategies were applied independently and their results were subsequently integrated (see ***Figure 5-1***). The following sections outline each strategy in more detail.

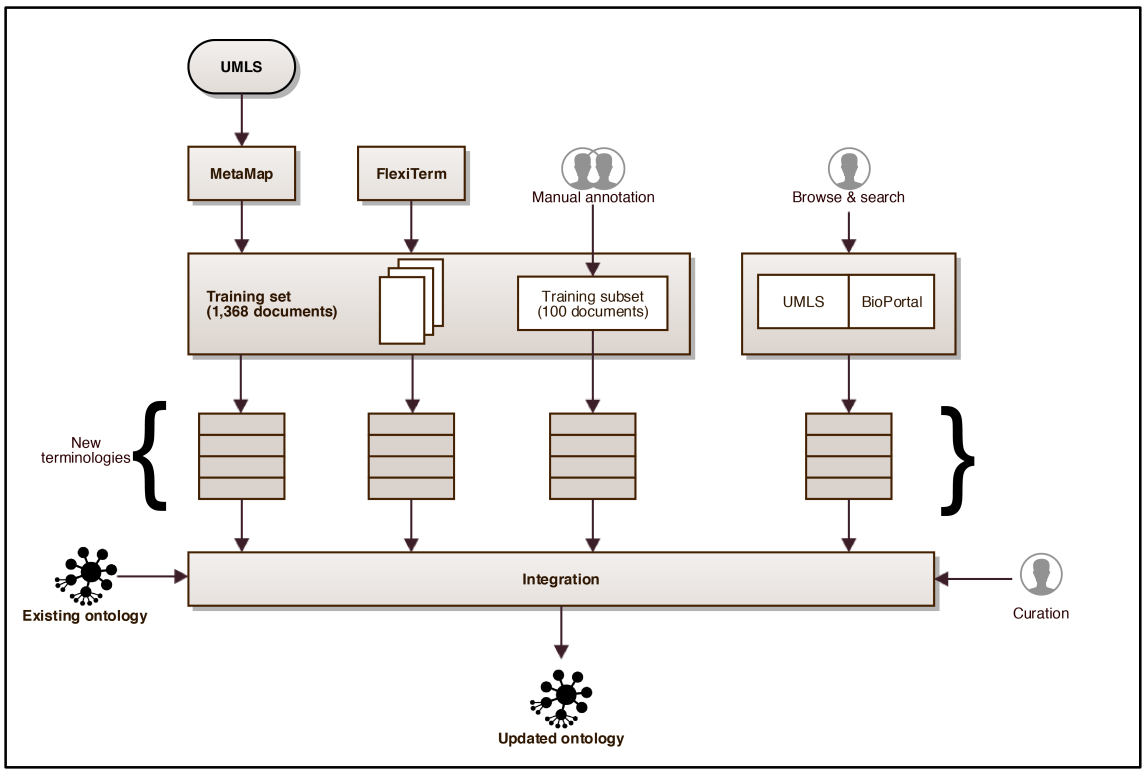


Figure 5-1 Ontology expansion strategies

5.1.1 Strategy 1: Dictionary-based term recognition

As previously mentioned in ***Section 4.4***, we considered both UMLS and RadLex for this process. Besides the fact that UMLS has better coverage than RadLex on radiology reports, we used MetaMap, a software tool dedicated for recognising UMLS concepts in biomedical text (Aronson, 2001). RadLex is not currently included as part of the UMLS Metathesaurus. Therefore, we decided to use UMLS instead of RadLex to search for potential terms.

Given the availability of MRI reports, we were now able to automate the process of finding other relevant concepts in the UMLS. We applied MetaMap on the training set to recognise UMLS concepts and obtain their unique concept identifier and a preferred name in the UMLS. Given that the majority of TRAK concepts were already cross–referenced to the UMLS, we used these identifiers to automatically remove known UMLS concepts from unnecessary consideration. The remaining MetaMap output formed a list of 1,121 UMLS concepts to be considered for possible inclusion in TRAK.

To facilitate the manual curation process, the list was ordered by the frequency of occurrence of each concept within the training dataset. The frequency graph shown in ***Figure 5-2*** shows a power law distribution (Clauset et al., 2009) of UMLS concept mentions. Using the Pareto principle (or 80:20 rule) as a guideline (Y. S. Chen et al., 1994), we focused on approximately 20% of most frequently mentioned concepts by considering only those that occurred at least 100 times in the training dataset. A total of 215 frequently mentioned UMLS concepts were manually curated and considered for inclusion in TRAK. Some examples of highest ranked relevant concepts include *intact, rupture, laceration*, etc.

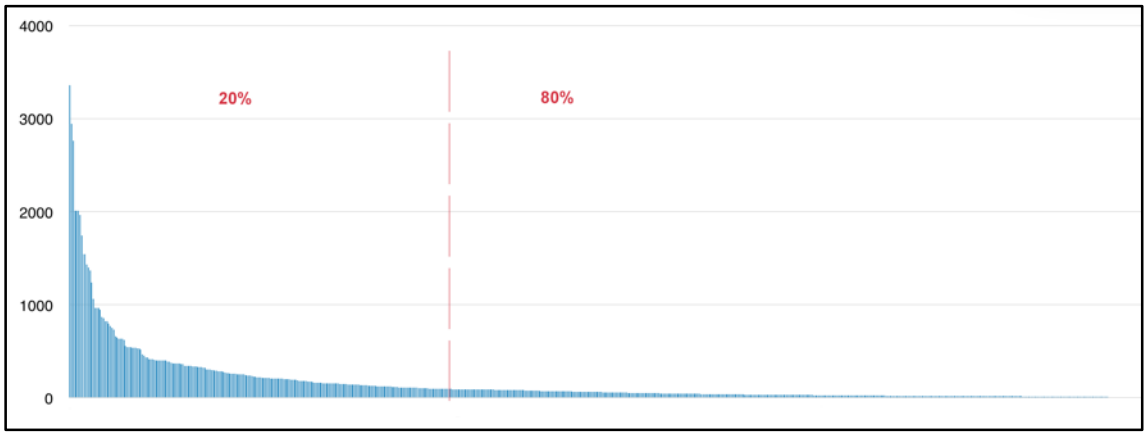


Figure 5-2 Power law distribution of UMLS concepts frequency to be included into TRAK from MRI reports

5.1.2 **Strategy 2: Automatic term recognition**

Using the UMLS to identify relevant concepts in text data has the advantage of providing not only their definitions and synonyms, but also their classification and a potential structure into which to embed them within the TRAK ontology. However, a previous lexical study conducted on a large corpus of various types of medical records (discharge summaries, radiology reports, progress notes, emergency room reports and letters) revealed that clinical narratives are characterised by a high degree of misspellings, abbreviations and non–standardised terminology (Hersh et al., 1997). The given study found that over 20% of the words used were unrecognisable, i.e. were not recognisable medical words, common words or names, and could not be algorithmically or contextually converted to such words. However, almost 78% of unrecognisable words were judged to be probably correctly spelled medical terms. These findings illustrate the challenges clinical narratives pose to dictionary–based term recognition methods such as that implemented by MetaMap.

In order to extract additional terms from the training dataset that were not found in the UMLS, we decided to enrich the original TRAK ontology into a lexicalised ontology by including these non-standard variants as new concepts or synonyms for existing concepts. Therefore, we complemented the use of MetaMap with FlexiTerm, an in–house data–driven method for automatic term recognition from a domain–specific corpus (Spasic et al., 2005).

FlexiTerm performs recognition of multi–word terms in three steps: linguistics filtering, normalisation and termhood calculation.

1. Linguistic filtering is used to select term candidates as follows. Once input documents have been pre-processed, term candidates are extracted by matching patterns that specify the syntactic structure of targeted noun phrases (NPs). These patterns are the parameters of the method and may be modified if needed. In our experiments, we used the following three patterns[1]:
   - `(JJ | NN)+ NN`, e.g. *anterior cruciate ligament*
   - `(NN | JJ)* NN POS (NN | JJ)* NN`, e.g. *Hoffa's fat pad*
   - `(NN | JJ)* NN IN (NN | JJ)* NN`, e.g. *medial condyle of tibia*

   Further, lexical information is used to improve boundary detection of term candidates by trimming leading and trailing stop words, which include common English words (e.g. *any*), but also frequent modifiers of biomedical terms (e.g. *small* in *small Baker's cyst*).
2. Term candidates are normalised in order to neutralise morphological and syntactic variation, linguistics phenomena commonly seen in biomedical terminology. The bag-of-word based normalisation process consists of the following steps:
   (1) Convert phrase into bag-of-words, and remove punctuation (e.g. ' in possessives), numbers and stop words including prepositions (e.g. of).
   (2) Remove any lowercase tokens with ≤2 characters.
   (3) Stem each remaining token.

[1] Explanation of tags and symbols: JJ (adjective), NN (noun, singular or mass), POS (possessive ending), IN (preposition or subordinating conjunction)

+: one or more occurrences, *: zero or more occurrences

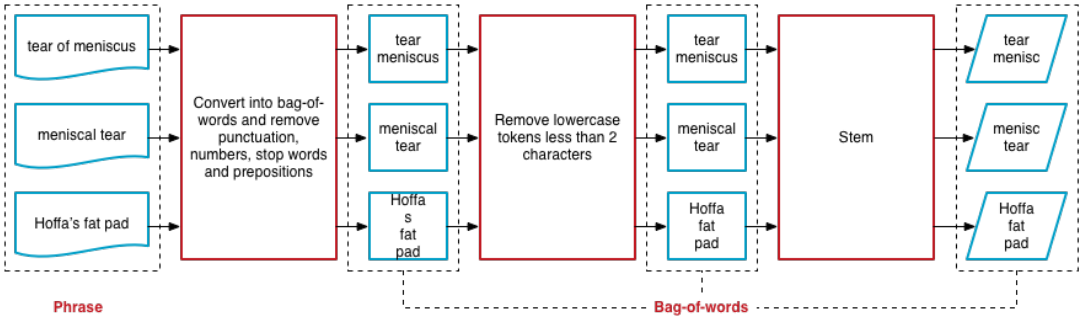


Figure 5-3 Example of the term candidate normalisation process with input *tear of meniscus*, *meniscal tear* and *Hoffa's fat pad*

For example, ***Figure 5-3*** shows a process that would map term candidates such as *tear of meniscus* and *meniscal tear* to the same normalised form {*menisc*, *tear*}, thus neutralising both morphological and syntactic variation resulting in two linguistic representations of the same medical concept. The normalised candidate is used to aggregate the relevant information associated with the original candidates, e.g. their frequency of occurrence.

While many types of morphological variation are effectively neutralised with stemming used as part of the normalisation process (e.g. *transplant* and *transplantation* will be reduced to the same stem), exact token matching will still fail to match synonyms that differ due to orthographic variation (e.g. *haemorrhage* and *hemorrhage* are stemmed to *haemorrhag* and *hemorrhag* respectively). On the other hand, such variations can be easily identified using approximate string matching. For example, the edit distance between the two stems is only 1 – a single insertion of the character a: *h[a]emorrhag*. In FlexiTerm, similar tokens are identified based on their phonetic and lexical similarity calculated with Jazzy (T. White, n.d.) (a spell checker API). Jazzy is based on edit distance (Damerau, 1964), but it also includes two more edit operations to swap adjacent characters and to change the case of a letter. Apart from string similarity, Jazzy supports phonetic matching with the Metaphone algorithm (Philips, 1990), which aims to match words that sound similar without necessarily being lexically similar.

Similar tokens are used to further normalise term candidates, and this makes FlexiTerm robust against orthographic variations in addition to morphological and syntactic variation and, therefore, suitable for use on clinical narratives.

3. Finally, termhood, a corpus–based measure that combines strength of collocation with frequency of occurrence, is calculated for normalised term candidates in order to rank them. The termhood calculation is based on the following formula:

Equation 5-1 Termhood value calculation (Spasić et al., 2013)

$$C-value(t) = \begin{cases} ln|t| \cdot f(t), & if\ S(t) = \emptyset \\ ln|t| \cdot (f(t) - \frac{1}{|S(t)|} \sum_{s \in S(t)} f(s)), & if\ S(t) \neq \emptyset \end{cases}$$

where T is a set of all candidate terms, t ∈ T, | t | is the number of words in t, f: T → N is the frequency function, P(T) is the power set of T, S: T → P(T) is a function that maps a candidate term to the set of all other candidate terms containing it as a subset. The method favours longer, more frequently and independently occurring term candidates.

In the original publication, FlexiTerm was thoroughly evaluated on five biomedical corpora including a subset of 100 MRI reports from the dataset used in this study. The highest values for precision (94.56%), recall (71.31%) and F-measure (81.31%) were achieved on this particular corpus.

Rank	Term variants	Score
1	mri knee mri knee mri mri let knee	991.4791
2	collateral ligaments collateral ligament collalteral ligaments collateral ligamets collateral ligmaments collateral ligments	912.8972
3	medial meniscus medial mensicus medial menisci	832.4333
4	lateral meniscus lateral menisci lateral mensicus lateral miniscus	778.9633
5	hyaline cartilage hyaline cartilages hyaline cartilege hyaline cartilge hyaline crtilage	611.6248
6	femoral condyles femoral condyle	477.2596
7	posterior horn posterior horns posteior horn poterior horn	473.1277

Figure 5-4 Part of FlexiTerm output on training set

Termhood values are used as evidence to select higher–ranked candidates as terms over the lower–ranked ones. FlexiTerm not only extracts terms from text, but it also groups term variants such as *infrapatellar fat pad*, *infra-patella fat pad* and *infra-patellar fat pad* together (see ***Figure 5-4*** for FlexiTerm output). Preferably, we would only include the

nominative singular form for nouns and the first person singular present indicative form for verbs into the ontology to preserve its strict formality. However, many of these variants are different lexical forms of a word, such as plural, adjective and verb forms of nouns, and different tense forms of verbs. For example, *meniscal* is the adjective form of *meniscus*, and *torn* is the past participle verb form of *tear*. Although we would not directly include these variants into the ontology, they could still be used as clues for identification of new concepts (e.g. *posterior horn* ranked seventh by FlexiTerm was added as a new concept in TRAK), but also identification of previously unknown names of existing concepts, which are easily mapped to a concept via its known names. For example, *lateral femoral condyle* was identified as a new synonym for a concept with identifier *TRAK:0001037* previously known only as *lateral condyle of femur*.

We ran FlexiTerm over the whole training dataset of 1,368 MRI reports and extracted 1,076 term candidates with a total of 1,422 term variants. To facilitate the manual curation process, the list of automatically extracted terms was ordered by their termhood calculated by FlexiTerm. The termhood graph shown in ***Figure 5-5*** shows a power law distribution. Therefore, relying on the Pareto principle, we focused on approximately 20% of highest ranked terms by considering only those with termhood over 20. A total of 222 automatically extracted terms were manually curated and considered for inclusion in TRAK.

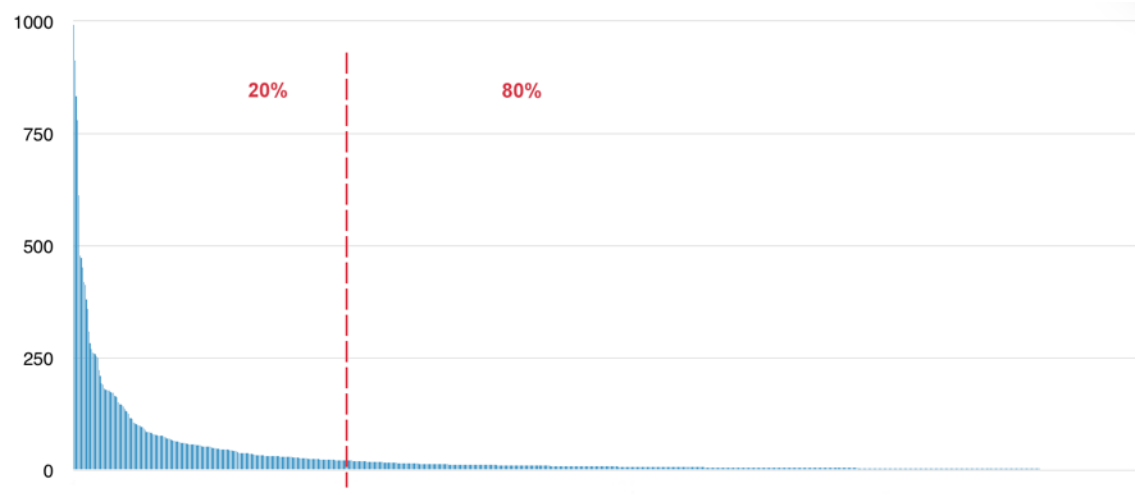


Figure 5-5 Power law distribution of FlexiTerm candidate termhood values

5.1.3 **Strategy 3: Manual data annotation**

As mentioned in Chapter 4, we manually annotated the development set and test set for the purposes of developing and testing the information extraction system. The annotated

test set was later used to create a gold standard for system evaluation. The annotated subset of 100 training documents was used to test the system during development, but also to inform the expansion of the ontology with terms manually annotated in text.

This strategy offers a potential to identify additional concepts and their names, particularly those that are non–standardised and occur less frequently in the training dataset. Recall that MetaMap identifies concepts based solely on the content of standardised medical dictionaries included in the UMLS. On the other hand, FlexiTerm may identify some non–standardised terminology, but in doing so it relies on the frequency of term occurrence. Moreover, FlexiTerm only extracts multi–word terms, thus ignoring concepts designated by a single word (e.g. *fissure, ganglion*, etc.). In addition to enabling us to identify relevant concepts overlooked by the previous two strategies, the annotation exercise allowed us to explore in detail how the terms were used in context, which helped disambiguate their meaning based on which they were embedded into the existing ontology structure.

Manual annotation tags were described earlier in ***Section 4.5***. These tags include:

- *finding* - clinical manifestations observed by a radiologist.
- *finding qualifier* - property of a finding
- *certainty* - certainty with which a radiologist diagnosed the given finding.
- *anatomy* - anatomical entity affected by the given finding.
- *anatomy qualifier* - further anatomical localisation
- *negation* - indication of a negative finding

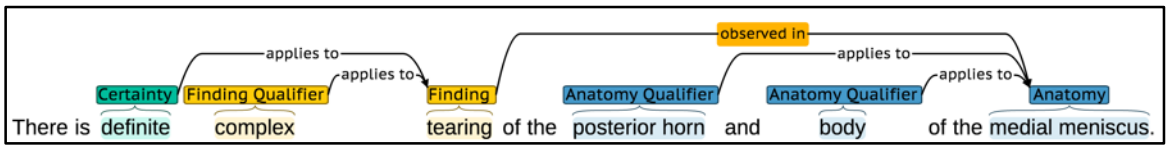


Figure 5-6 An example of manual annotation tags

Figure 5-6 provides an annotated example, where *tearing* is the finding, *definite* its certainty and *complex* its qualifier. The finding refers to *medial meniscus* as the affected anatomical entity, whereas qualifiers *posterior horn* and *body* provide further localisation of the anatomical site.

Table 5-1 Statistics of annotated terms on development set

Tags	Unique terms	Occurrences
finding	484	2,071
finding qualifier	113	284
certainty	68	202
anatomy	208	1,232
anatomy qualifier	178	469

Table 5-1 summarises the extent of the manual annotation on the development set. The fact that annotated terms were pre–classified into broad categories based on their labels allowed us to focus on particular branches of the TRAK hierarchy one at a time. In addition, some categories (e.g. *anatomy* and *anatomy qualifier*) were already extensively covered by the TRAK ontology. Therefore, the removal of 137 known terms referring to 60 TRAK concepts from unnecessary consideration greatly facilitated the manual curation and allowed us to consider all remaining phrases for potential inclusion in TRAK.

5.1.4 **Strategy 4: Manual terminology search**

So far, all three strategies for identification of new ontology concepts relied on the training dataset from which candidates were selected using a combination of automatic and manual methods. These data–driven approaches run a risk of overfitting the ontology based on the available data, which may result in incomplete coverage of the domain because some concepts (possibly the ones less frequently encountered in practice) were not mentioned in the available sample of MRI reports. In order to systematically cover the domain by including potentially relevant concepts that are not seen in the training dataset, we consulted two authoritative knowledge sources relevant for semantic interpretation of MRI reports: MEDCIN and RadLex.

MEDCIN

The first source, MEDCIN, was identified through the UMLS terminology services (UMLS, n.d.). MEDCIN is a medical terminology created and managed by Medicomp Systems, Inc., which contains more than 250,000 clinical concepts including symptoms, history, physical examination, tests, diagnoses and therapies structured into multiple clinical hierarchies. Magnetic resonance imaging of knee was found as one of those hierarchies. It provides a detailed taxonomy of findings that can be observed from knee MRI scans. We extracted this particular taxonomy from the UMLS by using MRI knee as a search term restricted to MEDCIN as the source vocabulary (see ***Figure 5-7*** for a screenshot of searching portal and results list).

All 703 concepts extracted from MEDCIN were named using phrases that represent detailed descriptions rather than traditional terms. As it can be seen from ***Figure 5-7***, these phrases were structured as follows: most concept names start with the same header (*magnetic resonance imaging of knee*) signifying that it belongs to a particular hierarchy in MEDCIN, followed by a detailed description of a finding. After removing the common header, *magnetic resonance imaging of knee*, from these phrases, we decomposed them into four categories: *finding, finding qualifier, anatomy* and *anatomy qualifier*.

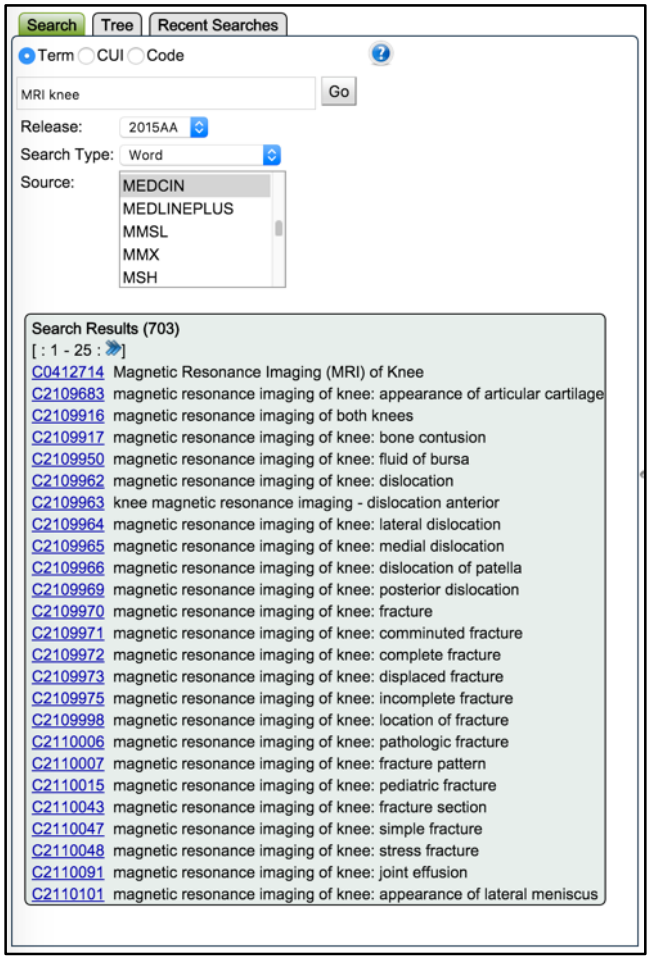


Figure 5-7 Searching MEDCIN with keyword *MRI knee*

For example, in the example taken from MEDCIN shown in ***Figure 5-8***, *osteochondral injury* represents the finding, *acute* its qualifier, *lateral femoral condyle* the anatomical entity affected, whereas *posterior aspect* is its qualifier, which provides more specific location for the given finding.

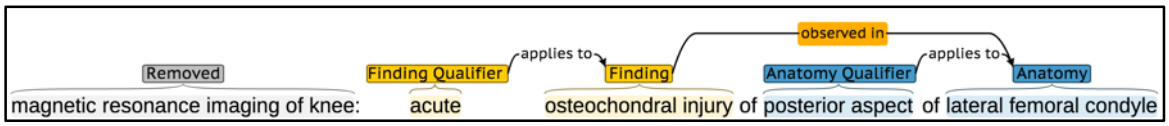


Figure 5-8 An example of decomposition of MEDCIN item

Manually decomposed concepts were compared against the existing TRAK concepts. Most anatomical concepts were already covered in TRAK, so the manual curation process focused mainly on concepts related to findings and their qualifiers. The resulting list consisted of 76 concepts, which were then manually curated and considered for inclusion in TRAK.

RadLex

The second source, RadLex, was identified through BioPortal, the most comprehensive repository of biomedical ontologies (NCBO, 2013). MRI is a technique used in radiology, a medical specialty whose concepts are formally described in the Radiology Lexicon (RadLex) – a controlled terminology designed as a single unified source of terms for radiology practice, education and research in an attempt to fill in the gaps in other medical terminology systems (Langlotz, 2006).

Figure 5-9 Radlex descriptor branch screenshot

RadLex is currently not distributed as a part of the UMLS. A study conducted on a corpus of 800 radiology reports that represented a mixture of imaging modalities including MRI revealed that out of 11,962 mentioned terms found in RadLex, 3,310 terms (i.e. almost 28%) could not be found in the UMLS (Yetisgen-Yildiz et al., 2011). These facts imply that much of the RadLex terminology would not be identified by MetaMap, used previously to identify UMLS terms. Previous study (Friedman, 1992) suggested that these missing terms in UMLS compared with RadLex are mainly modifier information, such as certainty, degree and change types. Therefore, we systematically explored RadLex using its distribution via BioPortal, especially focused on its *RadLex descriptor* branch (see ***Figure 5-9***). Leaf node children of this branch are mainly adjectives (rather than noun phrases, which is customary for terms) that can be used to describe radiology findings by specifying their qualifiers (e.g. *lobulated* would be a qualifier of a *cyst*). We consulted domain expert and considered a total of 41 subclasses out of which 13 were relevant for MRI reports (these classes are indicated with an asterisk in ***Figure 5-9***). These subclasses were used not only as the source of potential terms for TRAK, but also to provide a structure for incorporating such terms into TRAK. Out of a total of 439 terms, 167 terms were cross-referenced to RadLex. The *RadLex descriptor* class has been renamed to *finding descriptor* and embedded into TRAK as a subclass of *quality*.

5.2 Results

Each strategy suggested a list of terms to be considered for inclusion or exclusion. ***Table 5-2*** shows percentages of finally included terms out of suggested terms. These included terms contributed to the expansion of new concepts and synonyms.

Table 5-2 Contribution of each ontology development strategy towards final inclusion in to TRAK

Strategy	Suggested	Included	Percentage
Dictionary-based term recognition	215	29	13.5%
Automatic term recognition	222	77	34.7%
Manual data annotation	-	-137	-
MEDCIN search	703	76	10.8%
RadLex search	439	167	38.0%

Given that the majority of the original TRAK had already been cross-referenced with UMLS, the inclusion percentage of dictionary-based term recognition and terminology search from MEDCIN is relatively low, at 13.5% and 10.8% respectively. However, as these terms come from existing mature knowledge recourses, they also contributed to the final expansion with their corresponding synonyms and relationships.

Due to high prevalence of misspelled, abbreviated and non-standard varied terms, which are not included in standard knowledge resources, automatic term recognition resulted in a relatively higher inclusion rate of 34.7%. 38.0% of RadLex suggested terms were included, which is also a relatively high percentage. These terms were systematically added as complementary concepts in addition to those data-driven strategies because of limited training set size.

Although manual annotation did not suggested terms for inclusion, it helped to remove concepts already existed in TRAK from terms suggested by dictionary-based and automatic term recognition, which further improved the efficiency of manual curation process.

Domain knowledge support is consistently required across these strategies, especially for manual data annotation and terminology search from MEDCIN and RadLex. Domain knowledge support is also essential for manual curations.

As the final result, the original TRAK ontology was expanded from 1,292 concepts, 1,720 synonyms and 518 relationship instances to 1,621 concepts, 2,550 synonyms and 560 relationship instances.

Each of the four strategies made unique contribution to the final expansion of TRAK. Dictionary-based and automatic term recognition are primarily automated processes, and thus require little human effort and domain knowledge in recognition processes. However, manual data annotation, manual terminology search and final curation did require support from domain professionals. Sufficient time was also required to finish these processes and to ensure their qualities. Therefore, although the manual terminology search resulted in higher final inclusion percentages, it was time-costly and thus less efficient.

Chapter 6 KneeTex: A system for information extraction from knee MRI reports

In this chapter we describe a system we developed to extract information from knee MRI reports. The system itself represents the main contribution of this thesis to health informatics. To our best knowledge, no other system operates in this domain. The methodology we implemented in our approach represents more general contribution to computer science. While the idea of ontology-driven information extraction is by no means new, traditionally, the extent of knowledge engineering involved in the development of domain–specific ontologies with sufficient detail and coverage for text mining applications led them to be regarded as prohibitively expensive. In the previous chapter we described how the knowledge extracted from text using advanced NLP could be curated and used to rapidly update the content of biomedical ontologies. In this chapter, we demonstrate how such an ontology can serve as a fine–grained lexico–semantic knowledge base and play a pivotal role in guiding and constraining information extraction achieving results in line with human–like performance.

6.1 System specification

Information extraction (IE) is the task of automatically selecting specific facts about pre–specified types of entities and relationships from free–text documents. In other words, the goal of IE is to convert free text into a structured form by filling a template (a data structure with predefined slots) with the relevant information extracted (slot fillers) (Cowie and Lehnert, 1996). We derived this template from the annotation schema defined previously in ***Section 4.5***. These slots and restricted relationships among them also exactly match the annotation tag set and guidelines.

Figure 6-1 provides a graphical representation of a template specific to our system, whose structure is illustrated using Unified Modelling Language (UML) (Jacobson, 1999). The template specifies the types of entities and relationships we aim to extract in this particular study. The goal of our system is to extract clinical observations described in MRI reports. A clinical observation usually consists of two major parts: finding and anatomy. Finding is what is observed from MRI scans, e.g. an injury or a disease. Anatomy describes the location of the finding, usually a specific anatomical entity. Both finding and anatomy may come with additional information attached. Within the template, we refer to them as qualifiers.

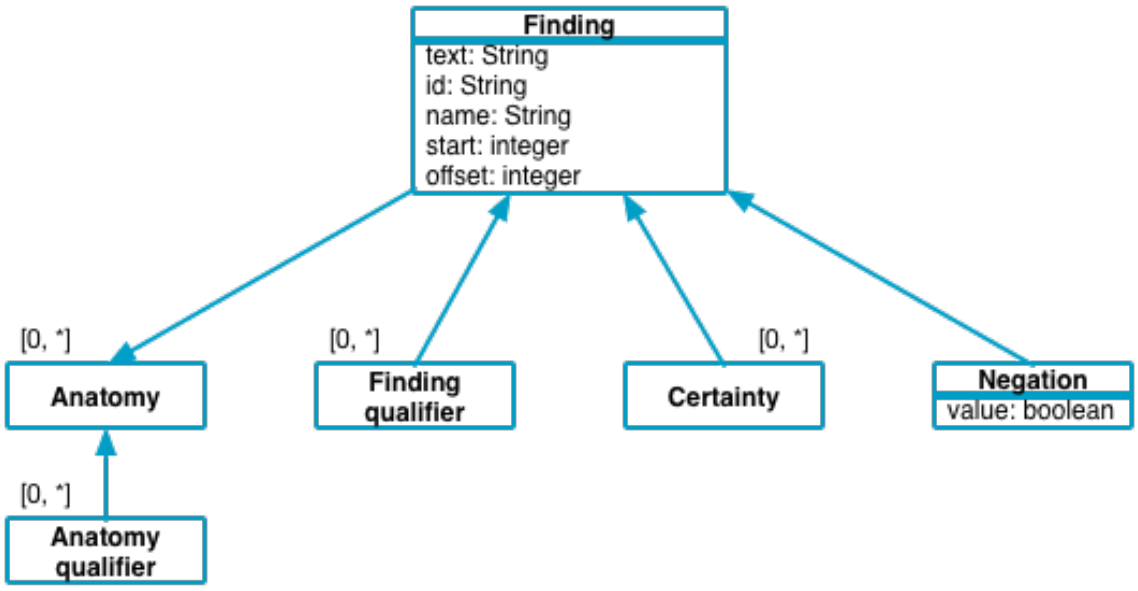


Figure 6-1 KneeTex information extraction template represented using UML diagram

Figure 6-2 provides an example of headwords and linked qualifiers. Together with anatomy, its qualifier provides more specific location information of the finding, e.g. *lateral* as an anatomy qualifier to *tibial plateau* indicates it is the *lateral aspect of tibial plateau*. Finding qualifier is more complex in the sense that it can modify the finding in different ways, e.g. size, shape, severity, direction, injury type, etc. For example, in ***Figure 6-2***, *fracture* is the finding and *small* indicates size of the *fracture*, *depression* indicates the type of the *fracture*.

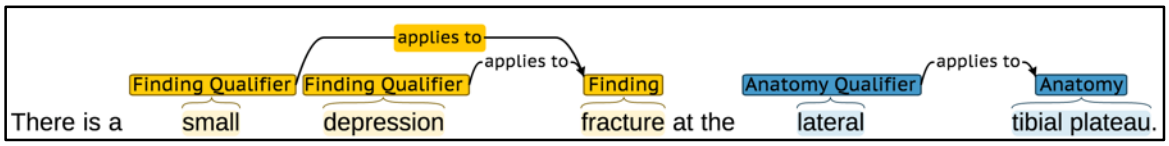


Figure 6-2 An example of headwords and qualifiers

For finding, there are two specific qualifier types in addition to general qualifiers: certainty and negation. Certainty qualifier describes the certainty level of related finding judged by the radiologist. Negation qualifier specifies whether described finding is actually found. See ***Figure 6-3*** for an example of certainty and negation qualifiers.

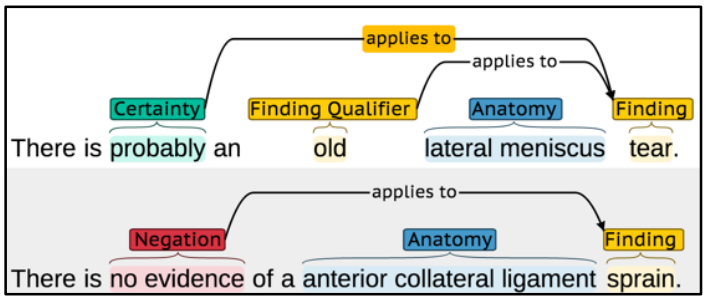


Figure 6-3 An example of certainty and negation qualifiers and how they are related to finding

In this template, each slot contains actual content extracted and its location in the document. The extracted information is further mapped onto the corresponding concept in the TRAK ontology. The concept's unique identifier and preferred name are also retrieved from the ontology and used to facilitate interpretation of extracted information.

Figure 6-4 and ***6-5*** provide examples of a filled template based on information extracted automatically from the given sentences.

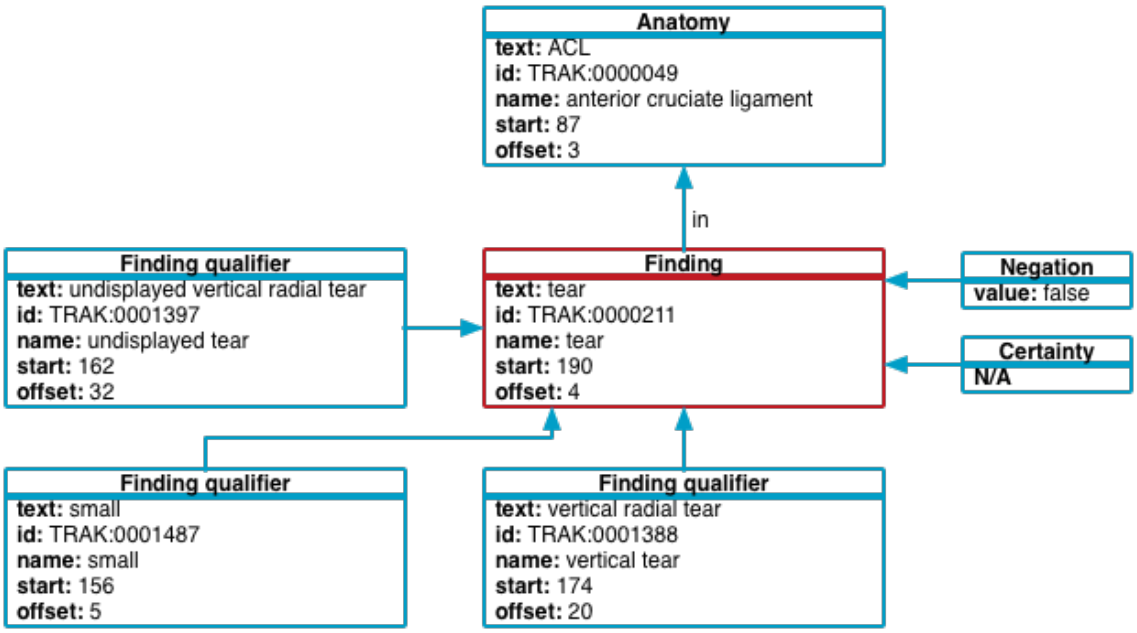


Figure 6-4 An example of a filled template from original text: *'There is a small undisplaced vertical radial tear of the posterior horn of the lateral meniscus.'*

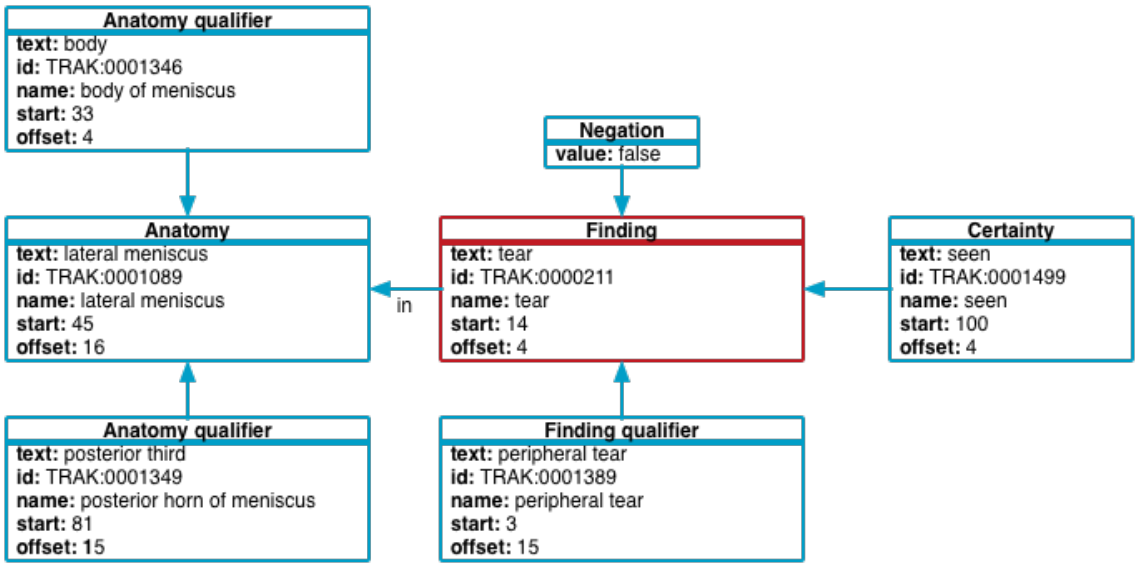


Figure 6-5 An example of a filled template from original text: *'A peripheral tear involving the body of the lateral meniscus extending into the posterior third is seen.'*

Once coded, the extracted information can be searched systematically. For instance, note that in the given examples equivalent phrases, *posterior horn* and *posterior third*, were mapped to the same concept, which allows for the extracted information to be searched by the underlying meaning and not merely its surface realisation in text. Note that KneeTex is an IE system and as such does not include an interface to search through the extracted information. However, the JSON format of extracted information allows for it to be stored directly into a document–oriented database such as MongoDB (mongoDB, 2015a), from which it can be easily queried.

6.2 System overview

Figure 6-6 shows the overall structure for KneeTex information extraction system. Modules shown above will be explained in the following sections.

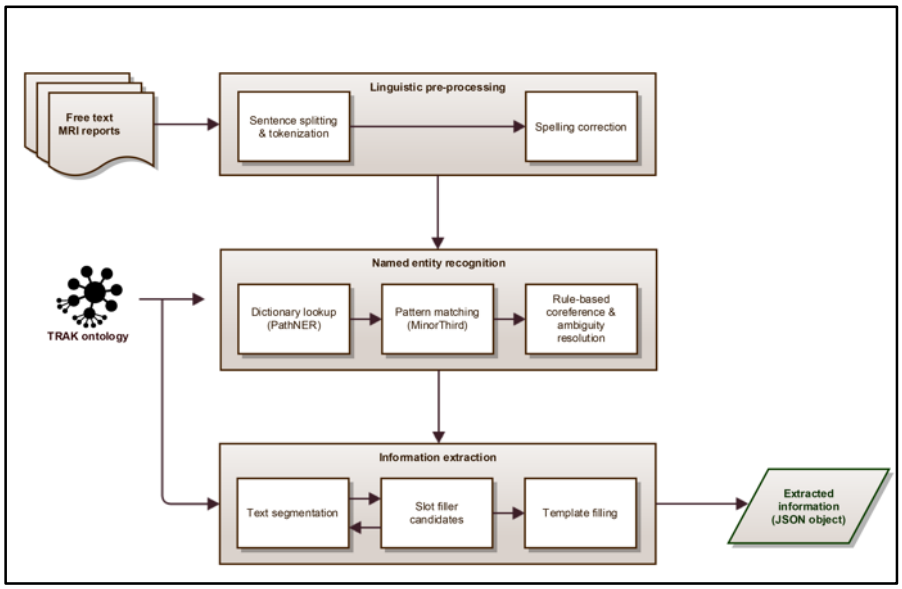


Figure 6-6 KneeTex system structure

6.3 Linguistic pre-processing

Previous lexical analysis on medical records shows that problems, such as sublanguage characteristics, misspellings and abbreviations, could pose difficulties for NLP applications (Hersh et al., 1997). Therefore, the training set needs to be prepared before more advanced NLP processes.

We pre-process the training set before applying further analysis, as in many other clinical NLP systems (Raja and Jonnalagadda, 2015). Most commonly used text pre-processing techniques include segmentation, tokenisation, spell checking and part-of-speech (POS) tagging.

The original MRI reports in dataset were all provided as one line documents. Thus they need to be segmented first. We used Stanford NLP for sentence segmentation and tokenization (Manning et al., 2014). In total, there were 13,051 sentences, 166,824 tokens and 3188 distinct tokens from the training set.

Levenshtein distance is used for typographical error recognition. A threshold value was set at 3 to avoid likely erroneous matching. By matching the training set with a comprehensive medical dictionary from Medline Plus (Merriam-Webster, n.d.) against the training set, we found that there were a total of 1,138 typographical errors from documents in the training set, including misspelling and merged words. On average, there were only 0.83 errors per document. With such low rate of typographical errors per document, it would not significantly affect process performances at later stage. Therefore,

we decided to choose a robust named entity recognition solution instead of implementing a separate spelling correction module.

Part-of-speech (POS) tagging is normally a component of the pre-processing stage in many NLP systems. However, Ferraro et al. noticed that leading POS taggers, including Stanford POS tagger (Toutanova et al., 2003; Toutanova and Manning, 2000), OpenNLP POS tagger (OpenNLP, 2010), LBJ POS tagger (Roth and Zelenko, 1998) and LingPipe POS tagger (Alias-i, 2008), have seen with performances declined by 8.5% to 15% in accuracy when tested on clinical narratives (Ferraro et al., 2013). Fan et al. (2011) also do not recommend directly applying pre-trained POS taggers on other datasets due to declined performance, even within the same domain. Meanwhile, compared to semantic tags, POS tags do not provide enough information to solve clinical problems (Taira, 2009). Considered that POS tagging is one of the first processes, to avoid cascading problems, we also have not included POS tagging at the pre-processing stage.

6.4 Dictionary lookup

Having sufficiently expanded the original TRAK ontology, its vocabulary can now be used to drive named entity recognition, whose aim is to automatically identify and classify words and phrases into predefined categories such as diseases, symptoms, anatomical entities, etc. In effect, NER is used here to identify candidates for slot fillers and as such represents the main vehicle of IE. The performance of dictionary–based NER approaches varies across different dictionaries and tools. A recent evaluation of three such state–of–the–art tools on a set of eight biomedical ontologies showed that their performance in terms of F–measure varied from 14% to 83% (Funk et al., 2014). ConceptMapper (a component of the Apache UIMA Sandbox (Ferrucci and Lally, 2004)) generally provided the best performance. Beside performance, we considered the ease of use. While converting an OBO ontology to ConceptMapper's dictionary format is straightforward, one must adopt the UIMA framework in order to use this particular component. For flexibility reasons, we opted to use PathNER (Wu et al., 2013) as an alternative to ConceptMapper.

PathNER (Pathway Named Entity Recognition) is a freely available tool originally developed for systematic identification of pathway mentions in the literature. On a pathway–specific gold–standard corpus, PathNER achieved F-measure of 84% (Funk et al., 2014). It implements soft dictionary matching by utilising the SoftTFIDF method (Cohen et al., 2003), a combination of the term frequency-inverse document frequency

(TF–IDF) (Salton and Buckley, 1988) and the Jaro-Winkler distance (Winkler, 1990). This makes the dictionary lookup robust with respect to the problem of term variation commonly seen in biomedical text, which often causes dictionary lookup based on exact string matching to fail (Tsuruoka et al., 2007). Typical term variations include morphological variation, where the transformation of the content words involves inflection (e.g. *lateral meniscus* vs. *lateral menisci*) or derivation (e.g. *meniscus tear* vs. *meniscal tear*), and syntactic variation, where the content words are preserved in their original form (e.g. *apex of patella* vs. *patella apex*) (Spasić et al., 2013).

In order to use PathNER to identify TRAK terms in text, we extracted the vocabulary from the ontology and converted it into PathNER's internal dictionary format. In effect, PathNER is used here to identify TRAK terms in text as candidates for slot fillers and as such represents the basis for template filling. We identified a few potential issues in the context of the given template. These relate in particular to the fact that the template requires two distinct main types of named entities: anatomical entities (e.g. *organ*, *tissue*, etc.) and findings (e.g. *injury*, *disease*, etc.). In order to systematically classify knee conditions, TRAK incorporates a knee–relevant portion of the Orchard Sports Injury Classification System (OSICS) Version 10 (Rae and Orchard, 2007).

Table 6-1 An excerpt of conversion from ontology vocabulary to PathNER dictionary

Ontology (OBO format)	PathNER dictionary
id: TRAK:0000513 name: ACL rupture xref: OSICS-10: KJAR id: TRAK:0000049 name: anterior cruciate ligament def: "A major stabilising ligament in the knee that attaches the surfaces of the femur and tibia." synonym: "ACL" EXACT [] id: TRAK:0000211 name: tear def: "Forcible tearing or disruption of tissue." synonym: "rupture" EXACT [] synonym: "tearing" EXACT [] synonym: "disruption" EXACT []	TRAK:0000513 ACL rupture TRAK:0000049 anterior cruciate ligament TRAK:0000049 ACL TRAK:0000211 tear TRAK:0000211 rupture TRAK:0000211 tearing TRAK:0000211 disruption

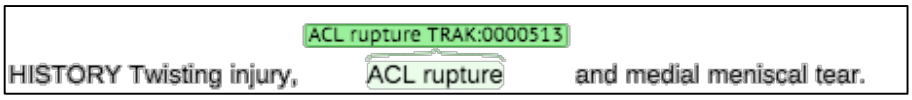
OSICS-10 is a classification system in which all classes encompass two types of information: (1) type of condition (injury or disease) and (2) anatomical entity affected by the condition. Our own approach to formal modelling of knee conditions was to

separate these two aspects and represent them by two distinct semantic types that correspond to finding and anatomy. For example, TRAK incorporates the following three terms: *ACL rupture*, *ACL* and *rupture* (see ***Table 6-1*** for details).

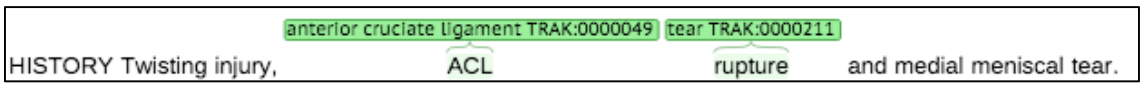
Obviously, the term *ACL rupture*, originally imported from OSICS–10, encompasses the other two terms. While the nature of taxonomic classification taken in OSICS–10 is useful for a range of applications in epidemiologic research (Rae and Orchard, 2007), it may pose problems for NER en route to template filling. Namely, PathNER looks for the longest possible match. This means that, given the three dictionary entries, the longest match in the following sentence:

HISTORY Twisting injury, ACL rupture and medial meniscal tear.

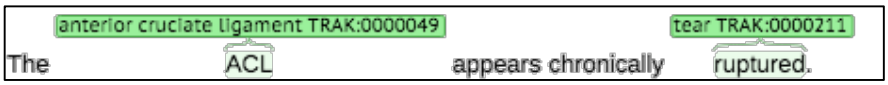
would result in the following annotation:



Alternatively, two separate annotations of *ACL* and *rupture* as follows:



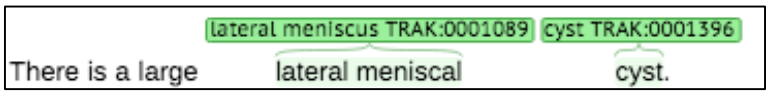
would greatly simplify the process of template filling, since the two recognised named entities can be mapped directly to the corresponding slots in the template (*anatomy* and *finding* respectively) based on their ancestries in the ontology (*anatomical entity* and *injury* respectively). The use of composite terms during NER could also give rise to inconsistent annotations, because sub–terms may occur wide apart in text, e.g.



At this point, we addressed two other problems associated with NER, namely ambiguity resolution and recognition of informal names. For example, we noticed that the term *joint effusion* (TRAK:0001410) defined in TRAK as *"Increased fluid in synovial cavity of a joint"* was commonly used in our dataset to refer to its child node *knee effusion* (TRAK:0001411). Although this can be solved later using hyponymy ambiguity resolutions, by safely assuming that in the context of knee MRI reports *joint effusion* will always refer to *knee effusion*, we ignored the concept identified by TRAK:0001410 and did not export it into PathNER's dictionary format. Instead, as a one-off solution, a

dictionary entry was created to map *joint effusion* to TRAK:0001411 instead in order for PathNER to recognise its intended meaning within the given context.

Further, we added some new entries to PathNER's dictionary in order to improve the performance of NER. The reason why such terms were not directly included into the ontology itself is the informal status of such terms (e.g. *tib–fib joint* is an informal synonym of *tibiofemoral joint*), and as such they do not belong to a controlled vocabulary. Given that it is customary for terms to be noun phrases (Justeson and Katz, 2008), we also limited the use of adjectives and verbs to the leaf nodes of the *finding descriptor* branch as explained earlier. Still, we needed to use these lexical classes as part of NER as we noticed from the training dataset that adjectives and verbs were commonly used to refer to concepts formally described in TRAK. For example, in the following sentence:



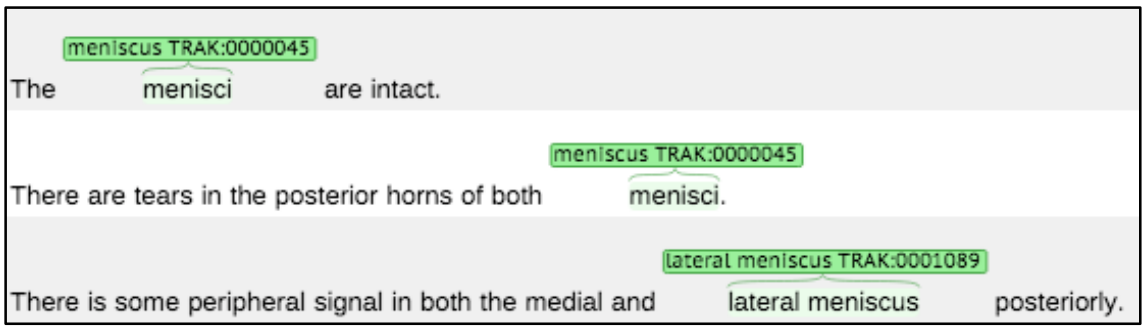
lateral meniscal refers to *lateral meniscus* (TRAK:0001089) in which the finding, i.e. *cyst* (TRAK:0001396), is noted. As previously mentioned in ***Section 5.1.2***, we preferably only included the nominative singular form for nouns and the first person singular present indicative form for verbs into the ontology to preserve strict formality. So it would be incorrect to specify *lateral meniscal* formally as an official synonym of *lateral meniscus* (TRAK:0001089) within the ontology. Therefore, instead, we encoded "unofficial" synonyms, which are not actually included in the TRAK ontology, separately within PathNER's dictionary, thus enabling the use of informal synonyms in NER while preserving the strict formality of the ontology. It was in this manner that the verb form *ruptured* was mapped to the term *rupture* (TRAK:0000211) in a previously discussed sentence. In total, the names of 128 concepts were ignored during ontology–to–dictionary conversion and 250 new entries were added to the dictionary.

6.5 Pattern matching

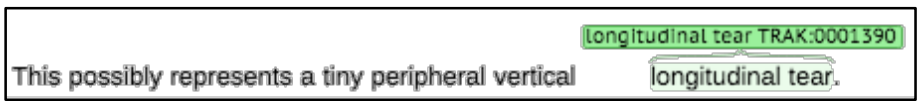
Previous dictionary lookup using PathNER does not recognise everything. There are still remaining concepts in the text need to be recognised. These concepts include those in nested phrases and negation triggers. To help segmenting documents into relevant sections, section headings will also need to be recognised. These remaining concepts and section headings frequently occur in the training set with regular occurrence patterns. Therefore, we implemented a few patterns in the Mixup pattern-matching language (Cohen, 2004) to address these remaining problems.

6.5.1 **Pattern-based named entity recognition**

Following the use of PathNER, a couple of NER–related problems may still persist. For example, consider the following three sentences:



with the terms *meniscus* (TRAK:0000045) and *lateral meniscus* (TRAK:0001089) recognised by PathNER. The analysis of their context reveals that all three references should actually be mapped to two concepts: *lateral meniscus* (TRAK:0001089) and *medial meniscus* (TRAK:0001090), both children of *meniscus* (TRAK:0000045). The first two annotations refer to an abstraction of two more specific concepts mentioned, which results in an ambiguous representation of the intended meaning. In the third sentence, correct identification of *medial meniscus* requires the coordinated expression *medial and lateral meniscus* to be interpreted as *medial meniscus and lateral meniscus*. Similarly, dictionary–based NER will fail to recognise enumerated terms. For example, the following sentence mentions three types of *tear* formally described in TRAK:



namely, *longitudinal tear* (TRAK:0001390), *vertical tear* (TRAK:0001388) and *peripheral tear* (TRAK:0001389), but only the rightmost one would be recognised by PathNER. Finally, in phrases such as *medial meniscectomy*, *patellar tendinitis* and *prepatellar bursitis*, PathNER will succeed in identifying terms referring to findings, i.e. *meniscectomy* (TRAK: 0001511), *tendinitis* (TRAK: 0000229) and *bursitis* (TRAK: 0000225), but it will not recognise implicit references to the anatomical entities affected, i.e. *medial meniscus* (TRAK: 0001090), *patellar tendon* (TRAK: 0000053) and *prepatellar bursa* (TRAK: 0001054).

In KneeTex, these linguistic phenomena are resolved using a set of 109 pattern–matching rules, whose results are used to correct or supplement annotations of named entities generated by PathNER. These rules were implemented in Mixup (My Information eXtraction and Understanding Package), a simple pattern-matching language (Cohen,

2004). For example, the following rules[2] illustrate the recognition of coordinated references to *medial meniscus*:

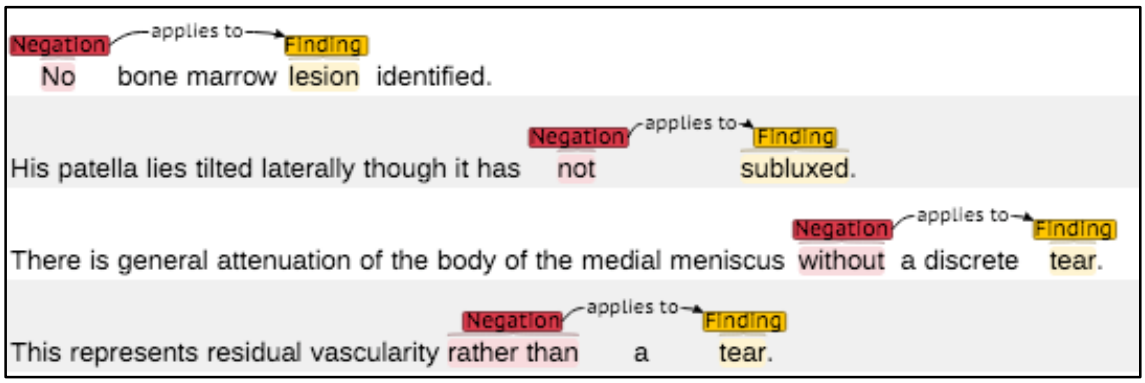
```
defSpanType conjunction =: ... [ eqi('and') ] ... ||
                           ... [ eqi('or')  ] ... ||
                           ... [ eqi('nor') ] ... ||
                           ... [ eqi('as') eqi('well') eqi('as') ] ... ;
defSpanType medialLateral =: ... [ eqi('medial') @conjunction eqi('the')?
eqi('lateral') ] ... ;
defSpanType TRAK:0001090 =: ... [ @medialLateral eqi('meniscus') ] ... ;
```

By applying PathNER to the development set that contains 100 randomly selected documents, it generated 4,439 annotations. In addition to PathNER generated annotations, 430 annotations were created using pattern-matching rules.

6.5.2 **Negation**

In addition to supporting NER, negation terms need to be identified to indicate the absence of explicitly mentioned findings. We initially considered using NegEx (Chapman et al., 2001a), a simple algorithm for identifying negated findings as mentioned in ***Section 2.3.5***, for negation term identification. Considering that the NegEx algorithm is also based on lexicons and regular expressions, we decided to build our own negation lexicons and patterns so that it could have the best adaption to the domain sublanguage.

We used the following terms as negation triggers: *no*, *not*, *without* and *rather than*. The following examples illustrate their use to negate findings:



[2] Mixup is a pattern language for text spans, i.e. token sequences. Keyword `defSpanType` defines a span type whose structure is specified to the right of the equal sign, where square brackets [and] indicate the start and end of a span respectively, `eqi('foo')` matches the token `foo` and ... matches any sequence of tokens. Postfix operator `?` specifies that the preceding token can be matched either once or not at all. Finally, operator || is used to specify alternative patterns.

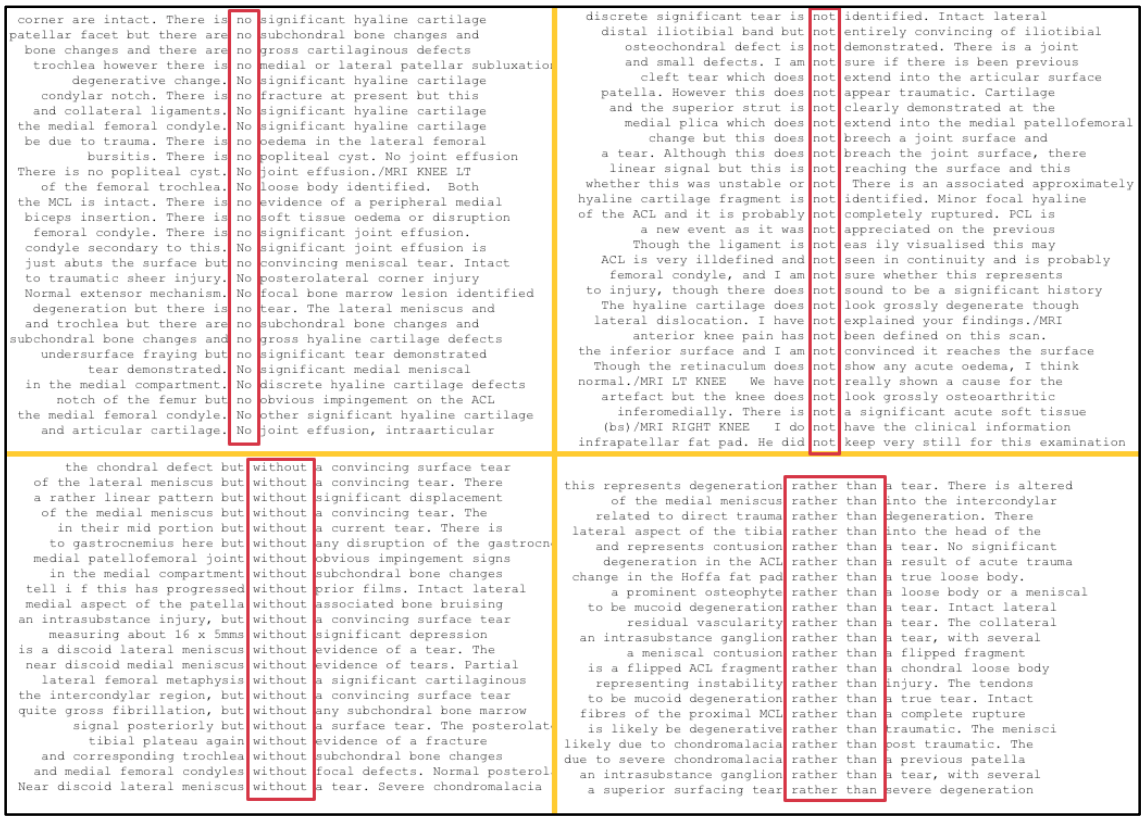
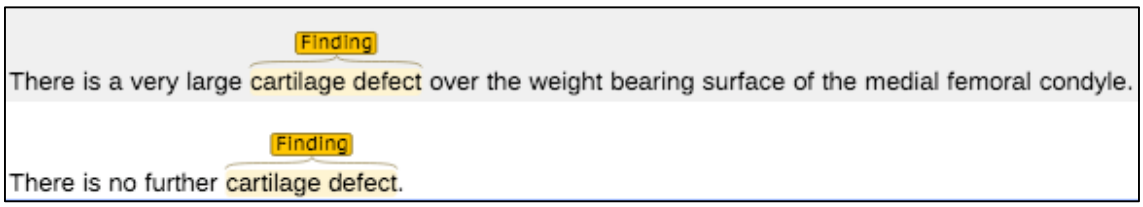


Figure 6-7 Concordances of negation terms

Based on our observations on the training data (see ***Figure 6-7***), all negation terms are assumed to occur before the finding they negate. We also defined a single exception to the negation rule. The negation term *no* is ignored when it is used as part of the phrase *no further*, in which case the finding is assumed to be positive, e.g.



6.5.3 **Section headings**

Although their structure varied across the data set, the given MRI reports generally tended to organise information under the following headings: *mri of the left/right knee*, *indication*, *history*, *findings* and *conclusion*. Their lexical and orthographic features were incorporated into a single pattern–matching rule designed to recognize a section heading as a sequence of upper case tokens from a list of fifteen.

6.6 Rule-based co-reference and ambiguity resolution

Once recognised, named entities are imported into a relational database and further scrubbed in order to disambiguate them. Semantic ambiguity may arise naturally from linguistic phenomena such as hyponymy, a relationship between a general term (hypernym) and its more specific instances (hyponyms), and polysemy, where a term may have multiple meanings. Multiple related interpretations may also arise from nested occurrences of named entities.

6.6.1 **Term Nestedness**

During dictionary lookup, PathNER will return longest possible matches with similarity scores over a certain threshold. As a result, there will be no overlap between named entities recognised in this manner. However, pattern matching used in the second phase of NER may introduce nested annotations of named entities. For example, in the coordinated expression *medial and lateral meniscus*, PathNER will recognise two terms from the TRAK ontology: *medial* (TRAK:0000031) and *lateral meniscus* (TRAK:0001089). Pattern matching will subsequently recognise a coordinated expression as a reference to *medial meniscus* (TRAK:0001090). The nested occurrence of *lateral meniscus* should be retained as a valid reference to a named entity. However, the nested occurrence of *medial* represents an unsuccessful match to another named entity, *medial meniscus*, and thus should be removed. The choice between retaining and removing nested occurrences of named entities is based on their semantic types. For example, all nested occurrences of terms descending from the concept *quality* (TRAK:0000133) defined as *"a dependent entity that inheres in a bearer by virtue of how the bearer is related to other entities"* are removed. This will remove nested occurrence of *medial* in the previous example, but also references to *radial* (TRAK:0001531) and *vertical* (TRAK:0000077) in the example shown in ***Figure 6-4***.

6.6.2 **Hyponymy**

Hyponymy is a lexical relationship between two terms, where one term (hyponym) is subordinate to the other (hypernym) (Stede, 2000). For example, *cruciate ligament* is a hyponym of *ligament*, and *complete tear* is a hyponym of *tear*. When a hypernym is mentioned, it could have multiple possible interpretations, either as the hypernym itself or as one of its hyponyms. Anatomical hypernyms cause higher level of ambiguity than finding hypernyms. ***Figure 6-8*** shows examples of hyponyms of *ligament* and *tear* in the TRAK ontology. Taking *ligament* as an example, it has 14 hyponyms in the TRAK

ontology. Therefore, when *ligament* is mentioned in text alone, it could have 15 different interpretations pointing to different anatomical locations. Although the mentioning of *tear* may have 16 different interpretations, these interpretations represent the same finding with different details. Meanwhile, finding hypernyms occur much less frequently than anatomical hypernyms. For example, the standalone mentioning of *ligament* occurs 82 times than 14 times of *tear* in the training set. Therefore, we focused on resolving anatomical hypernyms.

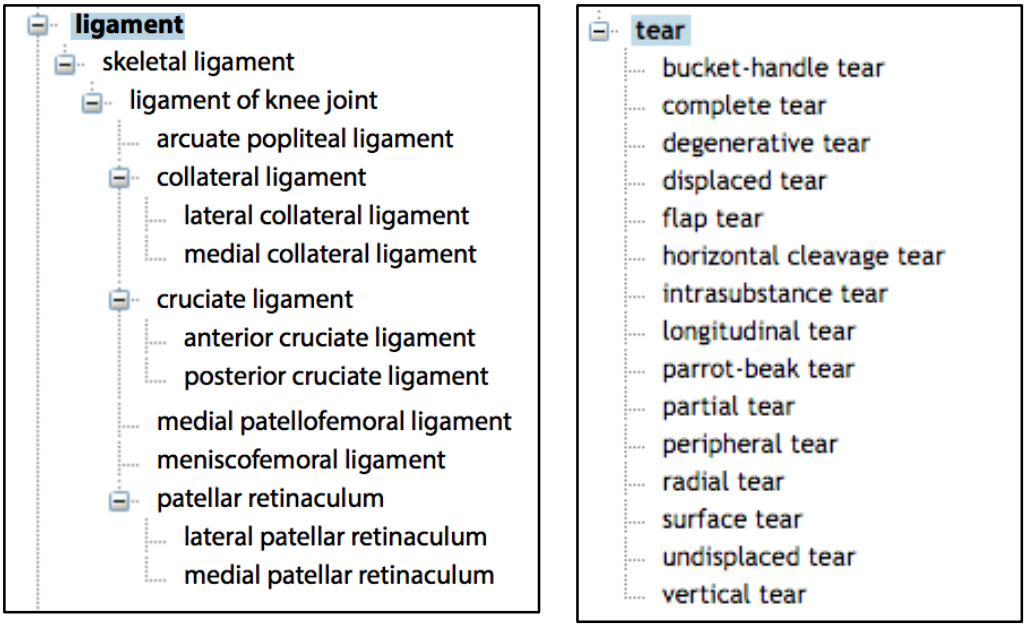
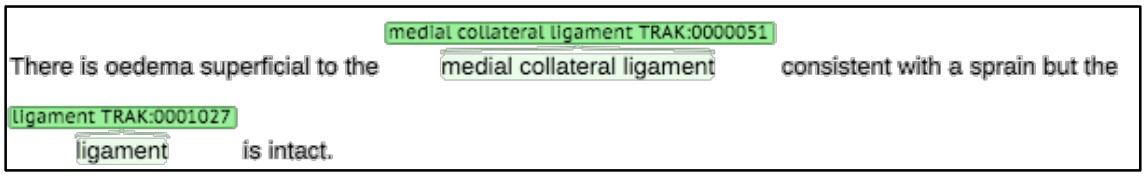
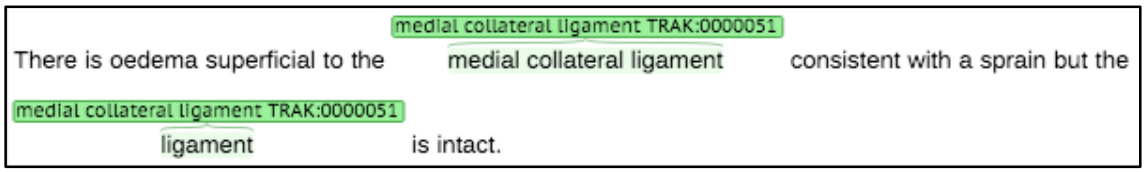


Figure 6-8 Examples of hyponyms of *ligament* and *tear* in the TRAK ontology

In clinical discourse, co-referential terms are often used to maintain text coherence and cohesion, e.g.



In this example, the hypernym *ligament* co-refers to its hyponym *medial collateral ligament*, and therefore its interpretation should coincide with that of the hyponym. In other words, the literal interpretation of the hypernym (*ligament*) obtained originally by dictionary lookup should be corrected using the annotation of the co-referring hyponym (*medial collateral ligament*), e.g.



This type of ambiguity is resolved systematically by identifying co-referential named entities, i.e. those that refer to the same concept. Co-reference resolution is applied to named entities recognised as one of the following concepts: *meniscus* (TRAK:0000045), *ligament* (TRAK:0001027), *tendon* (TRAK:0000046) or *muscle* (TRAK:0001088). In such cases, co-reference is resolved by looking for previous closest mentions of their ontological descendants. Once a hypernym term is spotted, the system looks for its ontological descendant term occurs before the start of current hypernym term. A distance threshold of 100 is set to avoid mapping to unrelated hyponym concepts. If multiple descendant concepts are found within the distance threshold, the closest one will be selected and used to replace existing mapping concept for current hypernym. However, if no descendant concept is found, the hypernym concept will remain as it is.

6.6.3 **Polysemy**

Polysemy refers to a linguistic phenomenon where a single word or a phrase may be associated with multiple meaning and therefore have the potential to be misinterpreted (Krovetz, 1997).

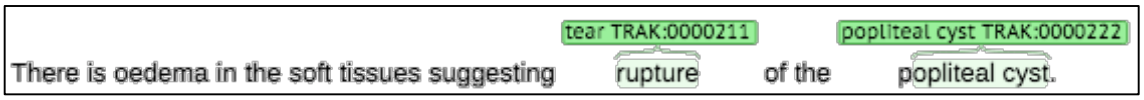
Sublanguages are restricted to specific semantic domains, which in turn affect the word usage. They generally tend to reduce the degree of polysemy. Nonetheless, the problem may still persist. For example, as pointed by the domain expert, word *rupture* in phrases *ligament rupture* and *cyst rupture* would be interpreted differently. In the former case it should be mapped to the following concept in the TRAK ontology:

```
id: TRAK:0000211
name: tear
def: "Forcible tearing or disruption of tissue." []
synonym: "rupture" EXACT []
synonym: "tearing" EXACT []
synonym: "disruption" EXACT []
synonym: "split" EXACT []
is_a: TRAK:0000206 ! injury
relationship: occurs_in TRAK:0000045 ! meniscus
relationship: occurs_in TRAK:0000046 ! tendon
relationship: occurs_in TRAK:0001072 ! skeletal muscle
relationship: occurs_in TRAK:0001027 ! ligament organ
```

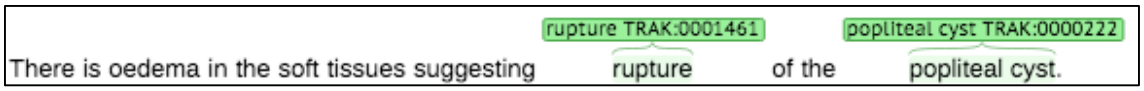
In the latter case, it should be mapped to an alternative interpretation represented by the following concept:

```
id: TRAK:0001461
name: rupture
def: "The result of breaking open or bursting." []
is_a: TRAK:0001456 ! morphologic descriptor
```

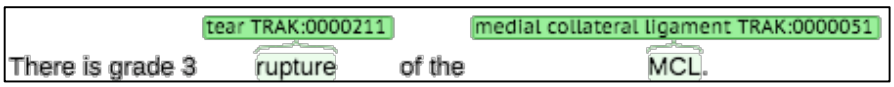
In such cases, co–occurrence information is used to resolve typical ambiguities observed in the training set. For example, when *rupture* co–occurs with a *cyst* (i.e. any descendant of the *cyst* concept), e.g.



it is used to correct its default interpretation as a *tear*, which represents an *injury*, to an alternative one, which represents a *morphologic descriptor*:



Thus, we are able to differentiate between the different uses of the term *rupture* in this latter example and that of the following example:



By default, *rupture* has already been mapped to the TRAK concept with identifier *TRAK:0000211* representing an *injury*. It co-occurs with *MCL*, the abbreviation of *medial collateral ligament*, which matches the relationship defined for concept *TRAK:0000211*.

This polysemy interpretation is restricted to the sublanguage of the knee injury domain. It may not be correct in or generalisable to other clinical domains.

6.7 Template filling

We previously described how the ontology, or more specifically – its vocabulary, is used to support NER as the first step in IE. Template filling as its final step is also driven by the ontology, or more specifically – its structure, i.e. relationships between concepts. This involves accessing information about semantic types by traversing the *is–a* hierarchy in order to identify slot filler candidates. In addition, relationships between the concepts are used to check compatibility between potential slot fillers. For example, if the extracted finding is a *tear*, then the anatomical entity affected must be *soft tissue* such as *ligament* or *tendon*. Similarly, if the affected anatomical entity is *cartilage*, then its qualifier must be related to *bone* or *joint*.

We originally considered using OntoCAT (Adamusiak et al., 2011) for this purpose, as it provides a programming interface to query ontologies shared on BioPortal or user–specified local OBO files. However, this would separate ontology querying from querying data, which are stored in a relational database. In order to enable integrative querying of both data and knowledge, we imported the ontology into the database. This allowed us to implement ontology–driven IE as a series of SQL queries that simultaneously access the data and the ontology. The remainder of this section describes the template filling process, where all semantic interpretations mentioned imply the use of such queries. To facilitate the slot filling procedure, a two-step text segmentation process is incorporated. The aim of this two-step segmentation is to split long and complex sentences into segments that contain at most one finding term in each segment. The first segmentation is purely based on lexical clues, while the second segmentation combines lexical clues with allocated slot types. Some lexical clues such as *and* and *with* sometimes are used to connect nested concepts (see ***Section 6.6.1***) and do not necessarily indicate a separate finding statement. Therefore, such lexical clues are not considered in the first segmentation stage. With identified *finding* terms, the second segmentation works on to further segment a sentence or a segment from previous segmentation when there are two *finding* terms separated by these lexical clues.

6.7.1 **Text segmentation**

In an effort to separate the contexts in which multiple findings are mentioned within the same sentence of an MRI report, these sentences are split into segments. Sentence segmentation involves separation of items in lists occurring within certain sections of MRI reports, namely those of history and conclusions. For example, the following two sentences

> HISTORY Twisting injury, tender medial joint line, positive McMurray's
> CONCLUSION Complex fragmented posterior lateral meniscal tear, complete ACL tear.

would be segmented into parts relying on comma as a separator. Other sentences are segmented using a set of lexical clues such as *but*, *which*, *consistent with*, etc. For example, the following sentence:

> There is oedema in the soft tissues at the posterolateral corner but the popliteal tendon is intact consistent with a sprain.

would be separated into three segments:

1. There is oedema in the soft tissues at the posterolateral corner
2. but the popliteal tendon is intact
3. consistent with a sprain.

Segmentation simplifies subsequent context analysis. When used in combination with the ontology to infer relationships between named entities, segmentation minimises the need for complex syntactic analysis. In fact, other than analysing prepositional phrases, no other syntactic analysis is performed as part of template filling in KneeTex. Alternatively, syntactic parsing can be used to support text segmentation, but such an approach would be more computationally intensive and not necessarily improving the accuracy. Due to the prevalence of ill formed sentences in clinical narratives (Fan et al., 2013; M. Jiang et al., 2015), lexical rules may be more robust. For example, in sentence '*He experienced decreased range of motion and tenderness'*, *decreased* should only link to *range of motion*. A person with specific domain knowledge will recognise this correctly. However, the parser incorrectly links *decreased* to both *range of motion* and *tenderness*.

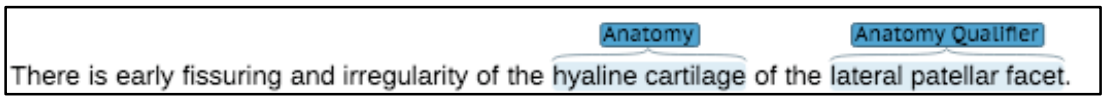
6.7.2 **Slot filler candidates**

Once the sentences have been segmented, previously recognised named entities are annotated as candidates for specific slots based on their semantic type. ***Table 6-2*** maps semantic types to the corresponding slots. For example, all named entities identified in the ontology as descendants of *certainty descriptor* (TRAK:0001422) or *visibility descriptor* (TRAK:0001495) are labelled as candidates for filling the *certainty* slot in the template shown in ***Figure 6-1***.

Table 6-2 Corresponding semantic types for slots

Slot	Semantic type	TRAK identifier	Example
Finding	accident	TRAK:0000362	Direct **fall** onto anterior tibia.
	clinical manifestation	TRAK:0000092	There is some **oedema** superficial to the MCL.
	modality-related characteristic	TRAK:0001447	The ACL returns **abnormal signal**.
	morphologic descriptor	TRAK:0001456	There is slight **thickening** of the medial collateral ligament.
	normality descriptor	TRAK:0001467	The articular cartilage is **unremarkable**.
	pathological condition	TRAK:0000204	There is a small **Baker's cyst**.
	physical examination	TRAK:0000656	Positive **McMurray's**.
	physiological condition descriptor	TRAK:0001482	No evidence of articular cartilage **damage**.
	surgery	TRAK:0000236	Presumably this had been excised during the ACL **reconstruction**.
Finding qualifier	clinical finding	TRAK:0000091	**Positive** McMurray's.
	composition descriptor	TRAK:0001322	Incidental note is made of a **simple** popliteal cyst.
	distribution pattern	TRAK:0001441	There is **focal** hyaline cartilage fissuring.
	orientation descriptor	TRAK:0001529	This could represent a **longitudinal** split.
	quantity descriptor	TRAK:0001468	There are also **several** loose bodies.
	size descriptor	TRAK:0001485	There is a **small** Baker's cyst.
	sport	TRAK:0000323	HISTORY **Squash** injury.
	stage of healing descriptor	TRAK:0001502	There is a **healing** tear of the medial collateral ligament.
	status descriptor	TRAK:0001478	Focal area of **severe** chondromalacia in the medial compartment.
	temporal descriptor	TRAK:0001488	There is **acute** ACL tear.
Certainty	certainty descriptor	TRAK:0001422	This raises the **possibility** of a previous patella dislocation.
	visibility descriptor	TRAK:0001495	Normal **appearance** of the articular cartilage.
Anatomy	anatomical entity	TRAK:0001337	The **menisci** , **collateral ligaments** and the **PCL** are intact.
Anatomy qualifier	anatomical location descriptor	TRAK:0001561	There is some oedema **superficial** to the MCL.
	general anatomical term	TRAK:0001581	There is a lot of oedema in the ACL **fibres**.
	meniscus zone	TRAK:0001345	Complex tear of **posterior horn** of the lateral meniscus.

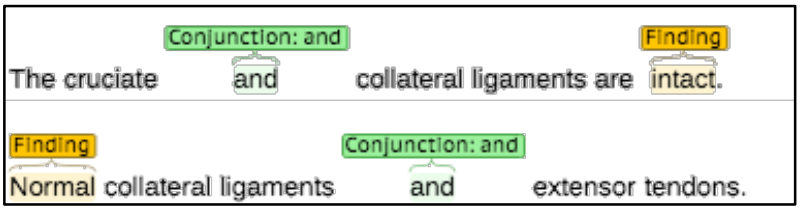
In addition, co–occurrence of certain concepts and semantic types is used to determine the most appropriate slot filler. For example, when *cartilage* co–occurs with other anatomical concepts, they are labelled as candidates for the *anatomy qualifier* slot rather than the *anatomy* slot as they otherwise would be, e.g.



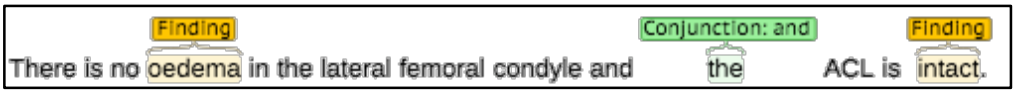
This is based on an observation that the *finding* will most likely apply to *cartilage* as an object, an observation drawn from the training data.

6.7.3 **Additional text segmentation**

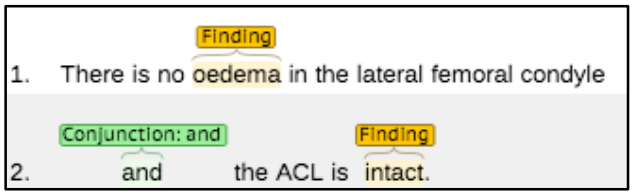
Preferably, we aim to have at most one finding in each sentence segment. However, there are still problems from the previous segmentation in ***Section 6.7.1*** using heuristics lexical clues. Some lexical clues such as *and* and *with* sometimes are used to connect nested concepts (see ***Section 6.6.1***) and do not necessarily indicate another statement about clinical findings, e.g.



Now with the *finding* slot candidates been identified, we can combine this information with such lexical clues in order to determine whether to use them to segment a sentence. For example, when two findings are separated by *and* as in:

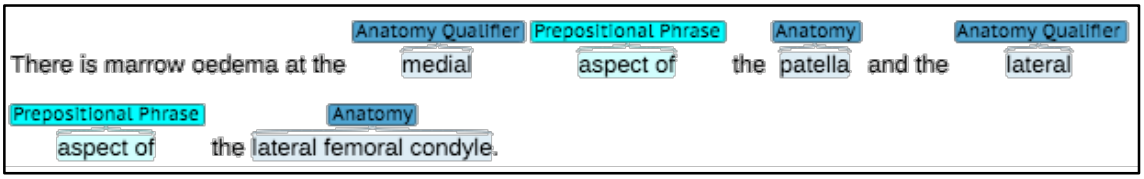


For the example shown above, once we have the two *finding* slot candidates identified, we could use the conjunction word *and* to split the sentence into:



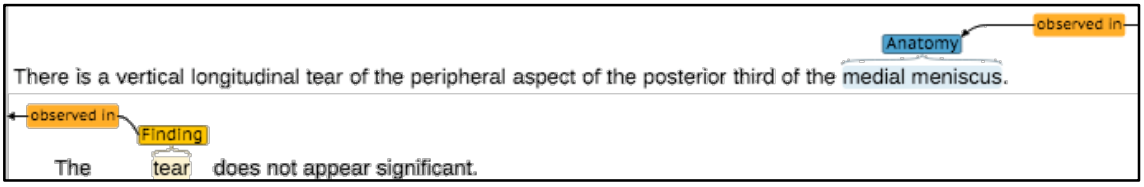
6.7.4 **Slot filling**

Finally, each segment is analysed in order to fill the template. In the first step, all *findings* are identified within a segment. Following the completion of the two–step segmentation process, most segments will contain at most two findings. The analysis of segments with a single *finding* involves identification of candidates for the following slots: *finding qualifiers, negation, certainty* and *anatomy*, which are all assumed to be linked directly to the given *finding*. Further analysis is required only if there are *anatomy qualifiers* that need to be linked to appropriate *anatomy* slot fillers. A simple analysis using lexical clues of prepositional phrases is used to achieve this. For example, in the following sentence:



the prepositional phrase *aspect of*, is used to link *anatomy* and *anatomy qualifier* slot fillers.

If no *anatomy* filler is found within a segment, an attempt is made to identify a potential filler within preceding segments. In the following example:



this approach would result in linking the mention of *tear* in the second sentence to *medial meniscus* mentioned in the previous sentence. In summary, when a single *finding* is found within a segment, the following workflow in ***Figure 6-9*** specifies the template filling rules.

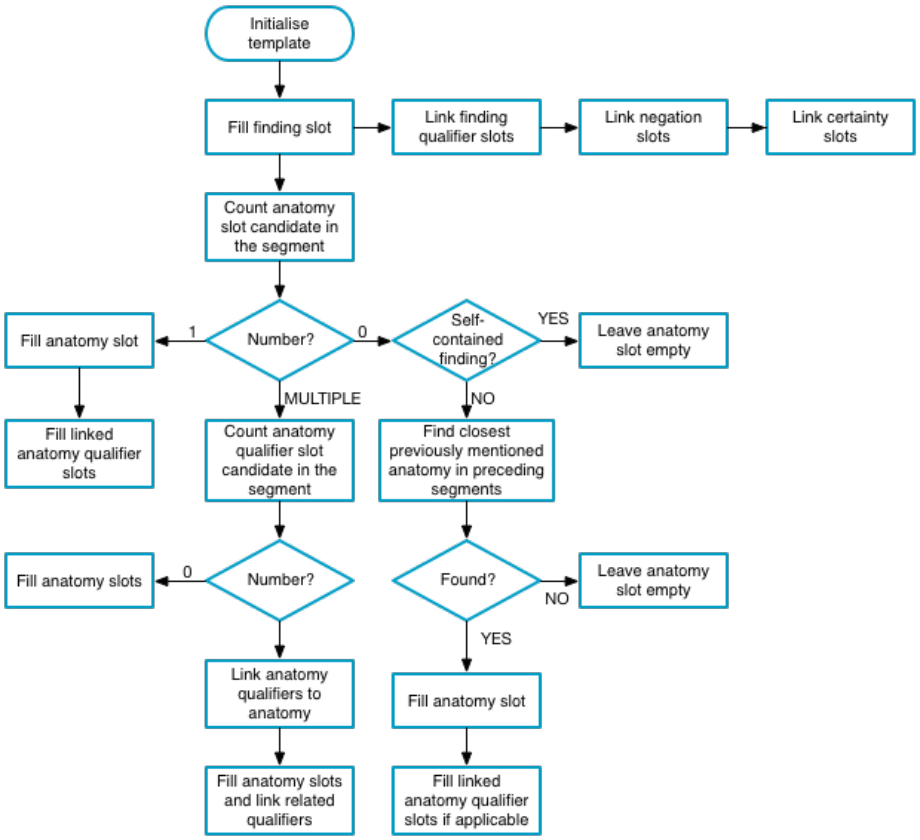
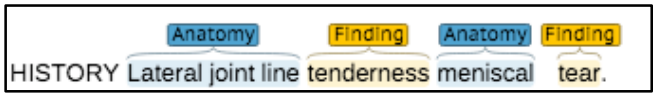


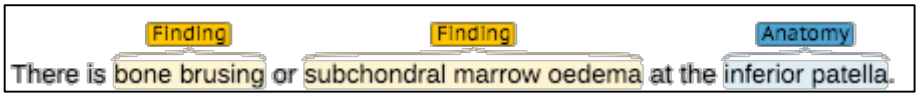
Figure 6-9 Template filling rule for segment with only one finding term

When two *findings* are identified within a segment, other slot fillers need to be linked to appropriate *finding*s. Using the end of the first *finding* as a boundary, the remaining slot

fillers are divided between the two *findings*. In the following example where a radiologist failed to enter a comma to separate two findings:



this approach would correctly link *tenderness* to *lateral joint line* and *tear* to *meniscus*. An exception is the use of conjunction *or*, e.g.



in which case the *anatomy* slot fillers are shared between the two *findings*. In summary, when two findings are found within a segment, the following workflow in ***Figure 6-10*** specifies the template filling rules.

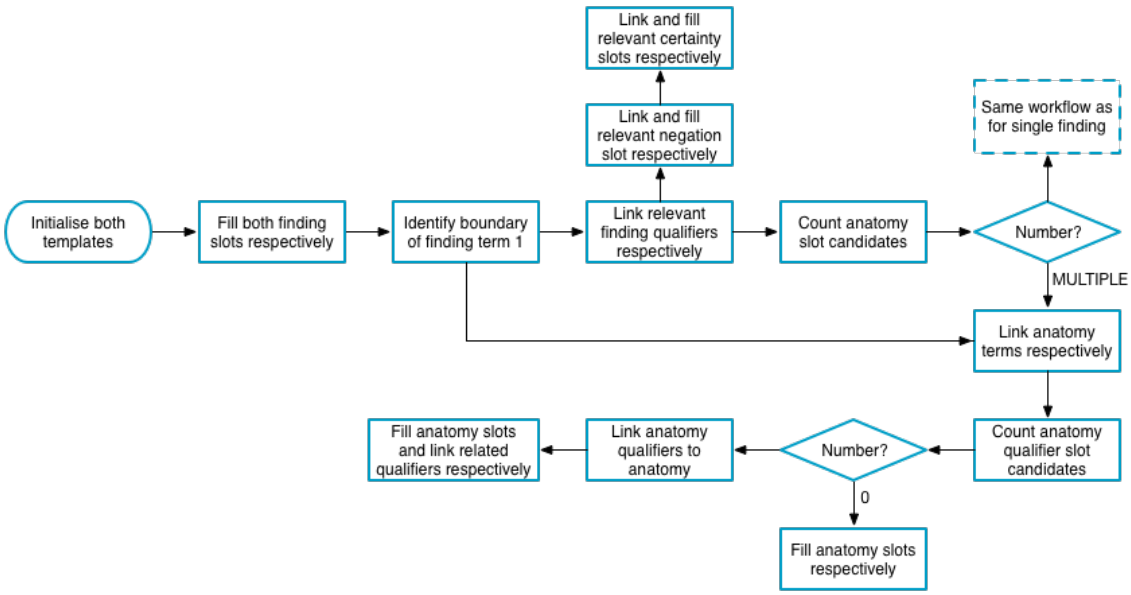


Figure 6-10 Template filling rule for segment with two finding terms

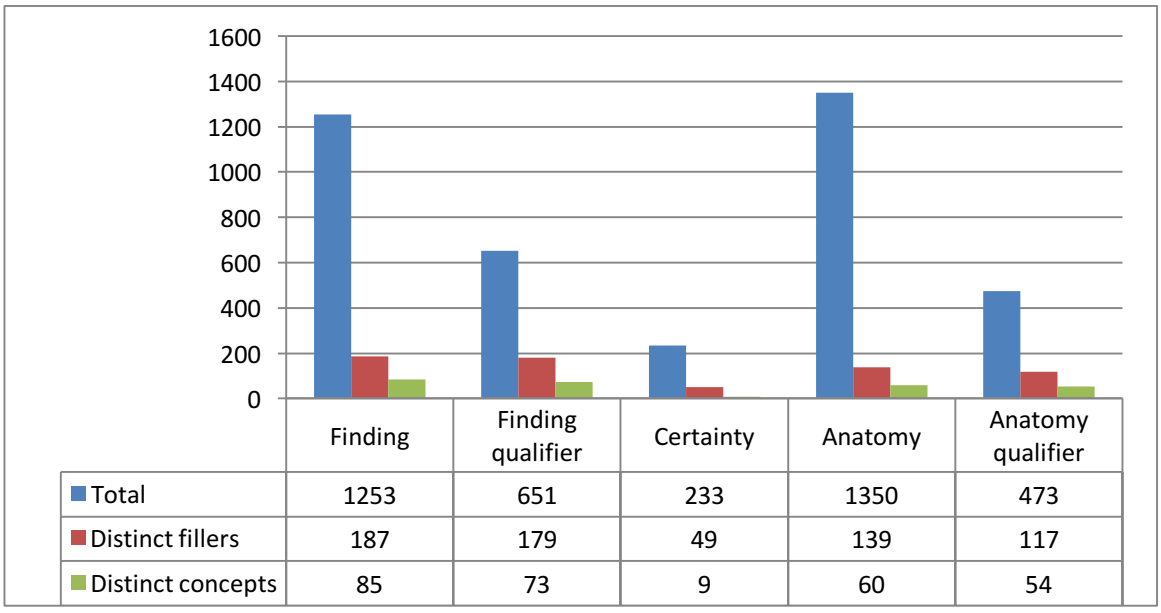
Although we made every effort to segment sentences, it is still possible to have more than two findings occurring in the same segment. Friedman (2006) pointed out rules that link those multiple findings together can be quite complex and it may be impossible for an NLP system to output extracted information without loss of substantial information. Therefore, when there are more than two findings occurring in the same segment, our system will not attempt to extract the segment. Instead, it will make a note to the segment for further inspections.

6.8 Results

6.8.1 **Gold standard**

A test dataset was created as a subset of 100 MRI reports selected randomly from the dataset described previously in ***Section 4.5.4*** and removed from consideration prior to system development. Its sole purpose was to test the performance of the system on unseen data. In order to create a gold standard, the test dataset was annotated manually by two independent annotators.

Inter-annotator agreement between the two annotators achieved 0.825 in Fleiss' Kappa coefficient value. This indicates almost perfect agreement according to the guideline (Landis and Koch, 1977) and therefore the annotation result was reliable. A gold standard was created by the third annotator who independently resolved the inter–annotator disagreements, ensured the consistency of annotations and mapped individual annotations of text spans to the corresponding concepts in the TRAK ontology. The gold standard annotations were converted to filled IE templates represented as JSON objects (see ***Textbox 6-1*** and ***6-2*** for examples) in order to support their comparison to KneeTex output during evaluation. ***Figure 6-11*** shows the distribution of slot fillers in the gold standard.



	Finding	Finding qualifier	Certainty	Anatomy	Anatomy qualifier
Total	1253	651	233	1350	473
Distinct fillers	187	179	49	139	117
Distinct concepts	85	73	9	60	54

Figure 6-11 Distribution of slot fillers in the gold standard

6.8.2 **Evaluation**

The system was evaluated at the conceptual level. Each automatically filled template slot was mapped to a TRAK concept. We compared manually annotated and automatically extracted mappings to TRAK concepts. In other words, each automatically filled template slot, represented by a TRAK concept identifier, was classified either as a true positive if it matched the slot filler in the gold standard or as a false positive otherwise. Conversely, each slot filler in the gold standard was classified as a false negative if it was not extracted by the system. Given the total numbers of true positives (TP), false positives (FP) and false negatives (FN), precision (P) and recall (R) were calculated as the following ratios:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The performance was evaluated using recall (R) and precision (P), as well as their combination into the F-measure:

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + ecall}$$

These values were micro–averaged across each slot (vertical evaluation) as well as the whole entries that take into account the links between the slot fillers (horizontal evaluation). ***Table 6-3*** provides evaluation results.

Table 6-3 System performances on test set over slots

Slot	TP	FP	FN	Precision	Recall	F-measure
Finding	1251	5	3	99.60	99.76	99.68
Finding qualifier	636	19	15	97.10	97.70	97.40
Negation	91	1	4	98.91	95.79	97.33
Certainty	232	8	2	96.67	99.15	97.89
Anatomy	1313	30	38	97.77	97.19	97.48
Anatomy qualifier	439	18	34	96.06	92.81	94.41
Overall	3962	81	96	98.00	97.63	97.81

6.8.3 **Stepwise performances**

The overall system performance achieved a satisfactory F-measure of 0.9781. Here we discuss contributions of different steps in the system.

We started with linguistic pre-processing. The Stanford NLP that we used successfully segmented 96 out of 100 documents. There were 4 documents not being segmented

correctly due to incorrectly use of the period punctuation. Therefore, the test set was segmented into 940 sentences compared with 942 in the gold standard.

Using dictionary converted from TRAK, we used PathNER for the dictionary lookup step and recognised 82.36% of all terms with F-measure of 0.79 for the original PathNER output compared with the gold standard. Next we applied complementary pattern-based NER and recognised 92.55% of all term, and improved the F-measure to 0.84. Further with rule-based ambiguity resolutions, 99.57% of all terms were recognised and the F-measure was also improved to 0.96. With pattern-based NER, 95.8% of all negations were recognised, achieving an F-measure of 0.97. Finally, 98.73% of all section headings were also recognised.

To evaluate the performance of co-reference resolution, we manually searched the test set to examine some concepts that occurs most frequently in the training set, including *meniscus*, *ligament*, and *tendon* that occur alone without qualifiers. 29 out of 32 occurrences of these concepts were mapped correctly to co-referenced concepts.

We also looked for occurrences of *rupture* in the test set to examine the performance of our polysemy resolution. All 6 mentions of *rupture* co-occurred together with different ligaments were correctly interpreted as *tear*.

In order to see how well the two text segmentation processes work, we manually examined segmented results on sentences that are longer than 200 characters. There are 21 of those sentences in the test set, and 20 of them require segmentation. Both segmentation processes worked well on these sentences and segmented them into 66 segments. However, there was also one unnecessary segmentation (see ***Figure 6-12***), though it would not affect the system performance:

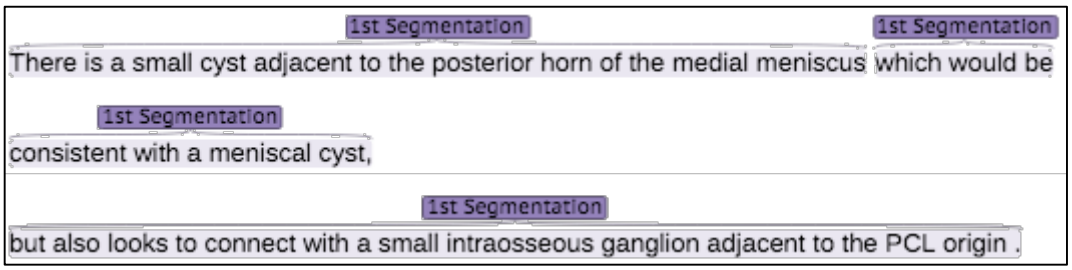


Figure 6-12 An example of unnecessary segmentation (the second segment: *which would be*) in the first text segmentation process

6.9 Discussion

6.9.1 **Performance comparison**

Given that most IE systems represent bespoke software solutions built around a task–specific template, we were not able to compare the performance of our system to that of an external system used as a baseline. However, we provide a like–for–like comparison in order to demonstrate that the performance of KneeTex is in line with the state-of-the-art in similar domains, albeit we admit that KneeTex operates in a relatively narrow domain. Given the fact that the systems operate in different domains and were tested on different datasets, we emphasise that any values cited hereafter are used purely as a reference point rather than a direct comparison.

We previously discussed two NLP systems that were applied to extract information from radiology reports, MedLEE (Hripcsak et al., 2002) and MPLUS (Christensen et al., 2002), whose recall was found to be 81% and 87% respectively. In KneeTex, recall ranged from 92.81% to 99.76% for individual slots, achieving 97.63% overall. With the overall precision of 98.00%, which varied from 96.06% to 99.60%, KneeTex compares well to MPLUS, which achieved 85% (Christensen et al., 2002).

Focusing on the negation slot, we compared KneeTex to other systems in terms of their performance in determining whether a given finding is negated. In particular, it is useful to compare KneeTex to NegEx, a widely used algorithm for identifying negative findings in clinical narratives, originally evaluated on discharge summaries where it achieved precision of 84.5% and recall of 77.8%. NegEx was incorporated into cTAKES,a generic NLP system tailored to the clinical domain, which identified negation in electronic medical records with F–measure of 96% (Savova et al., 2010). ConText, a derivative of the NegEx algorithm, was evaluated across different types of clinical narratives including radiology reports where it achieved precision of 100% and recall of 86%, thus yielding an F–measure of 93% (Harkema et al., 2009). KneeTex achieved slightly lower precision (98.91%), but balanced it with high recall (95.79%), giving an F–measure of 97.33%.

By viewing the process of filling the remaining slots as NER problem, we considered ConceptMapper, which recently demonstrated the best performance in terms of F–measure (83%) (Funk et al., 2014). In a different study, ConceptMapper was evaluated specifically for two types of entities: diagnoses and anatomical sites. Different configuration parameters were used in the experiments and the average values of precision, recall and F–measure were 86%, 90% and 88% for diagnoses and 91%, 88%

and 89% for anatomical sites (Tanenblatt et al., 2010). KneeTex recorded the following values for the F–measure: *finding* (99.68%), *finding qualifier* (97.40%), *anatomy* (97.48%) and *anatomy qualifier* (94.41%). As a matter of fact, PathNER, which drives NER in KneeTex, performed much better on our gold standard than on the pathway–specific corpus, when it was originally evaluated and reached 84% on F–measure, owing to a richer dictionary and additional post–processing, which included coordination, enumeration and co–reference resolution.

Overall, high recall and precision values, which are equivalent to human–like performance, can be attributed to two factors related to the fact that KneeTex operates in a relatively narrow domain with the document type restricted to MRI reports and their content confined to knee pathology. Firstly, the sublanguage used in this domain is proved to have consistent patterns, a property that makes it amenable to modelling by a set of sophisticated lexico–semantic rules. Secondly, the TRAK ontology in KneeTex provides a fine–grained lexico–semantic knowledge base, which is highly attuned to this sublanguage. Traditionally, the extent of knowledge engineering involved in the development of domain–specific ontologies with sufficient detail and coverage for text mining applications led them to be regarded as prohibitively expensive. However, previously we suggested that the knowledge extracted from text using advanced NLP could be curated and used to rapidly update the content of biomedical ontologies (Spasic et al., 2005). In this study we have demonstrated how this approach can be used in practice. Given that the principal link between text and an ontology is a terminology, which aims to map concepts to terms, we have focused on ATR as the most relevant NLP task in this context and shown how MetaMap (Aronson, 2001) and FlexiTerm (Spasić et al., 2013) can be applied for this purpose.

6.9.2 **Error analysis**

In order to get additional insight into the system's performance, we conducted error analysis and classified their major causes.

Dictionary lookup

Some of the errors stem from incorrectly recognised named entities. For example, in segment "*his patella tends to lie tilted laterally,*" string similarity caused PathNER to incorrectly recognise *patella tends* as *patellar tendon*, therefore failing to extract *patella* instead. While flexible matching based on a combination of the TF–IDF measure and the Jaro-Winkler distance makes this dictionary lookup robust with respect to the problem of

term variation commonly seen in biomedical text, it can obviously lead to incorrect matches based on string similarity. The right balance between flexible and incorrect matching requires careful experiments to optimise the similarity threshold used in PathNER.

As mentioned in ***Section 6.3***, POS tagging was not involved in the linguistic pre-processing due to declined performances on clinical narratives of leading POS taggers. They also provided less information than semantic tags that we have used. However, a properly domain-specific trained POS tagger could be considered to solve these remaining errors. A properly trained POS tagger adapted to this specific domain could be considered.

Co-references

In ***Section 6.6.2*** we described a simple heuristic approach we use to recognise co-reference. It is by no means a generic approach to co-reference resolution and, while our approach generally provides satisfactory results, some of the co-references will remain unresolved. In total, there are 3 co-references not being solved in the test set. For example, the following sentence:

> There is a cleavage tear of the lateral meniscus at the junction of the body and posterior horn which extends through the body but there is currently no evidence of a significant meniscal cyst.

would be segmented as:

> 1. There is a cleavage tear of the lateral meniscus at the junction of the body and posterior horn which extends through the body
> 2. but there is currently no evidence of a significant meniscal cyst.

The following anatomy slot filler was extracted from the second segment:

```
{"text":"meniscal", "id":"TRAK_0000045", "name":"meniscus"}
```

However, by looking at the whole, it is clear the term *meniscal* co-refers to the previous mention of *lateral meniscus*. Therefore, the correct slot filler should be:

```
{"text":"meniscal", "id":"TRAK_0001089", "name":"lateral meniscus"}
```

Additional experiments are needed to see how a generic co-reference resolution method would affect the system performance. While our approach may have lower recall, a more generic method would most likely have lower precision, so such method would not necessarily improve the system's performance. Nonetheless, it would improve the

generalisability of the system and facilitate its portability. Such investigation has been identified as an area of future work.

Preferred interpretation

Some of the errors were counted as such due to mismatch to manual annotations even though the extracted information can still be considered semantically correct. Consider for example the following sentence:

There is some blunting of the inner edge of the mid portion of the medial meniscus.

The system recognised the word *mid* as a synonym for *middle* (TRAK:0001598). Its classification as an *anatomical location descriptor* (TRAK:0001561) was used to fill the *anatomy qualifier* slot as follows:

```
{"text":"mid", "id":"TRAK:0001598", "name":"middle"}
```

Its literal interpretation as *"an intermediate part or section; an area that is approximately central within some larger region"* happens to be semantically correct. However, in the given context, the following annotation in the gold standard represents preferred interpretation:

```
{"text":"mid portion", "id":"TRAK:0001346", "name":"body of meniscus"}
```

since *body of meniscus* (TRAK:0001346) most specifically represents its middle third. Such mistakes happened 4 times in the test set. Although this may not affect understandings from the domain expert, the preferred interpretation would provide more precise information. Pattern-based rules could be implemented in the future to solve preferred interpretation problems with co-occurrence patterns.

Negation

Most negated representations used typical negation terms as *no*, *not*, *without* and *rather than*. However, negated representations using atypical negation terms that have negative meaning may be ignored. Consider for example the following sentence:

The low signal of the anteromedial bundle seen in a normal ACL is completely absent.

The system automatically extracted the following two findings:

```
{"text":"low signal", "id":"TRAK:0001309", "name":"low signal intensity"}
{"text":"normal", "id":"TRAK:0001312", "name":"normal"}
```

However, the system failed to make use of the clue *absent* found at the end of the sentence to recognise that these findings are actually negative. Although this mistake only happened once on the test set, it will be used to inform future improvements of the system.

6.9.3 **Generalisability**

Stepwise generalisability

Although we come up with a reusable system structure that can be referenced for other similar tasks, the availability of an ontology or other knowledge base or a dictionary is a prerequisite to use this system.

We used off-the-shelf tools including Stanford NLP and PathNER for linguistic pre-processing and dictionary lookup processes. Such processes can be directly reused without human interference, providing that there is an available knowledge source that can be converted into PathNER dictionary.

For the rest of the system, we used a lot of lexico-semantic rules to help with disambiguation and template filling. In general, it is not recommended to apply these rules directly on domains other than knee injury MRI reports. However, there are some thoughts that could be referenced to solve similar problems.

In our dataset, we found that radiologists quite often use nested descriptions such as *medial and lateral compartments*, probably for convenience. It is rational to assume that this may also happen in other clinical descriptions.

Hyponym and polysemy are common features of natural languages (Sheeba et al., 2013). Although the rules we used are restricted to the domain of knee injury MRI reports, such approaches of combining lexical and semantic patterns could be referenced to identify rules from targeted domain.

Slot candidates were annotated based on semantic rules and co-occurrence patterns. The TRAK ontology together with professional domain knowledge constructed these rules, i.e. these rules are domain specific. Similarly, the template filling rules are also domain specific.

Overall generalisability

Recall that our information extraction is ontology-driven, where the ontology used for this purpose was expanded by applying four strategies, three of which were data-driven. The data-driven strategies were applied against the training set, which does not overlap

with the test set. Nonetheless, both training and test datasets come from the same source upon which our ontology is dependant. Therefore, the results shown in ***Table 6-3*** may not be representative of the system performance on a different dataset where unknown concepts may be mentioned signifying incompleteness of the ontology.

In order to assess the generalisability of the system we conducted a series of stage–wise experiments in which we removed new concepts identified from the training dataset by using the three data-driven strategies. We specifically focused on concepts outside of the *finding descriptor* class for two reasons. Firstly, this class corresponds to the *RadLex descriptor* branch of the RadLex hierarchy and its dependency on the training data is minimal. Secondly, concepts from this class are used to fill three "leaf" slots (*finding qualifier*, *anatomy qualifier* and *certainty*, see ***Table 6-2***) that have no further dependencies (see ***Figure 6-1***) and as such will have no ripple effect on the template filling unlike *finding* and *anatomy* slots. For example, if *finding* is not identified, it will affect text segmentation as well as linking to other slot fillers. Therefore, the highest impact on evaluation results would be caused by concepts outside the *finding descriptor* branch.

Having identified just over 100 of such concepts, we randomly selected 100 of them, randomized their order and removed top k of these concepts (k = 10, 20, ... , 100) from the ontology, which was then used to run KneeTex on the gold standard. ***Figure 6-13*** provides a comparison of evaluation results. As expected, completeness of the ontology directly affected the recall of the system. This was most obvious when frequently referenced concepts such as *body of meniscus* (TRAK:0001346) or *joint effusion* (TRAK:0001411) were removed. However, the frequency and meaning of these concepts imply that they are of general relevance to the domain and not the result of overfitting to the training dataset. On the other side, the removal of less frequently referenced concepts did not have a profound effect on recall. For example, after removing as many as 50 concepts from the ontology, recall was still very high at 92.07% dropping by 5.57 percent points. Precision proved to be more stable reaching 94.48% after removing all 100 concepts, dropping only by 3.52 percent points. In summary, we can conclude that the system would most likely maintain high performance across different datasets. Ideally, we would like to test this assumption on such datasets, but at this point strict privacy laws prevent us from obtaining them from other institutions.

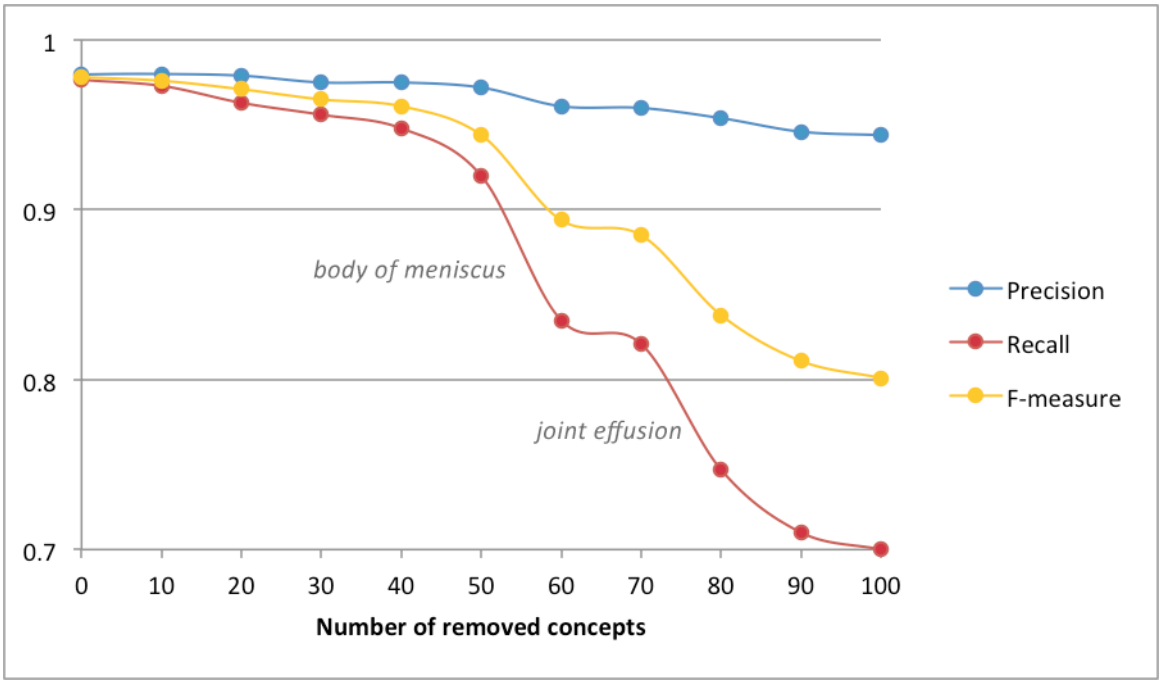


Figure 6-13 Stagewise experiments on system generalisability by removing concepts identified from training data

Chapter 7 KneeBase: a reusable web-based information retrieval system for epidemiologic study

In the previous chapter we described KneeTex, a system we developed for information extraction from knee MRI reports. Given an MRI report as input, the system outputs the corresponding clinical findings in the form of JSON objects. The extracted information is also mapped onto the TRAK ontology. As a result, formally structured and coded information allows for complex searches to be conducted efficiently over the original MRI reports, thereby effectively supporting epidemiologic studies of knee conditions. Note that KneeTex is an information extraction system and as such does not include an interface to search through the extracted information. In this chapter we describe KneeBase, an information retrieval system that supports this functionality, which represents an interdisciplinary contribution of this thesis.

7.1 Motivation

Ontologies and extracted clinical information can be used in many ways, such as predicting treatment outcomes, helping clinicians identifying potential related findings, answering clinical questions, etc.

Schattner et al. (2010) noticed improved clinical practice with support from extracted clinical information, even with technical failure in identifying some data. Westbrook et al. (2005) found that although experienced clinicians do not often use online information retrieval systems, but 88% of system users reported significant helpful for answering clinical problems. Overall, Buntin et al. (2011) concluded that 92% of recent health information technology related articles have indicated positive impact on healthcare delivery. Meanwhile, Bernard et al. (2012) have also noticed an increasing trend of using online information in younger practitioners and practices with Internet access.

Therefore, with demonstrated positive impacts described above, we developed KneeBase as an example of web-based information retrieval system that integrates the TRAK ontology and extracted information from KneeTex. We also provide a reusable system framework that can be used for other similar tasks.

7.2 System overview

We provided KneeBase as an internal demonstration of a clinical information retrieval system, which is one possible integrated use of ontology and ontology-based extracted information. A brief system structure of KneeBase is shown in ***Figure 7-1***. The system

is driven by previously extracted information from KneeTex and the TRAK ontology. System users are allowed to browse and perform direct or complex searches to obtain demanded information.

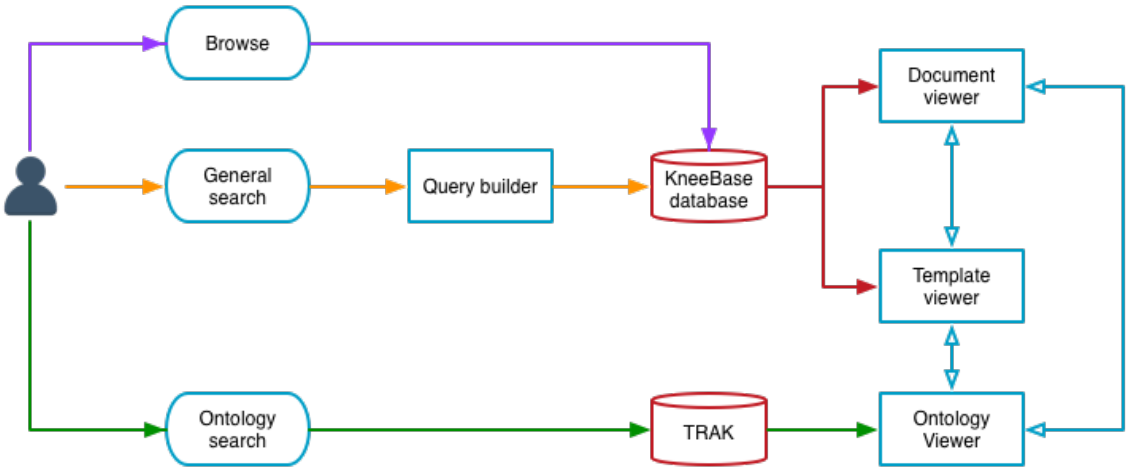


Figure 7-1 KneeBase system structure

7.2.1 **Structured data management**

KneeTex effectively converts unstructured text data into structured data representation that conforms to a predefined data model in the JSON format. The JSON format of extracted information allows for it to be stored directly into a document–oriented database such as MongoDB (mongoDB, 2015a), from which it can be easily queried.

Database structure

MongoDB stores data as BSON documents, which is a binary representation of JSON with additional type information. All documents are stored in collections, where a collection is a group of related documents that share common indexes (mongoDB, 2015b).

Table 7-1 KneeBase database structure

Collection	Content
documents	Original MRI reports in the development set.
sentences	Segmented sentences from reports.
sections	Extracted section headings.
terms	Terms mapping with TRAK ontology.
trak_concepts	Concepts from TRAK ontology.
trak_isas	Parent-child relations in TRAK.
trak_relations	Relations in TRAK other than parent-child relation.
templates	Filled templates.

Further permission will be required from NHS for practical use of this system. Therefore, for the system function demonstration purpose, we processed only 100 MRI reports with

KneeTex, which resulted in filling 1,375 templates represented as JSON objects. We stored their BSON equivalents into a MongoDB database. These BSON documents form the *templates* collection in KneeBase database. In order to support complex searches as part of epidemiological studies, we also imported seven other collections into the same database, see ***Table 7-1*** for details.

In order to support query expansion as part of information retrieval, KneeBase also integrates the TRAK ontology. The TRAK ontology is integrated in two ways: the ontology itself and derived database collections. Three collections derived from the ontology are: *trak_concepts*, *trak_isas* and *trak_relations*. These collections correspond to the vocabulary, taxonomy and a network of domain-specific relationships respectively.

7.2.2 **Core functionality**

Developed as a web-based information retrieval system, KneeBase allows cross-platform usages. ***Figure 7-1*** depicts a basic system working flow, which allows users to browse and search for both integrated ontology and extracted information. User views including document viewer, term viewer and ontology viewer are cross-linked to allow cross-search from returned search results.

The ontology is viewable either as tree structure enabled by using NCBO Ontology Tree Widget (NCBO, 2015) to visualize the ontology, or as list of terms from the database.

System users are able to search the database using customised queries, achieved by implementing the jQuery QueryBuilder (Sorel, 2015). Queries can be generated by using pre-defined fields (e.g. *category*, *filename*, *finding, finding qualifier, concept name*, *negated*, etc.) with user-defined keywords.

7.2.3 **Example use cases**

Main use cases of KneeBase will be query centred uses. Here we illustrated a few use cases.

Single condition query

1. User provide login information
2. System grant access if correct login information provided
3. User use query builder to submit a simple query, e.g.if the user wants to search for information related to ***tear***:
   - Choose search field: **Term**
   - Fill keyword field: ***tear***
4. System returns result in term viewer, provide term id, definition, etc.

5. Optional further research:
    a. User request to see the term in ontology hierarchy
    b. request to see documents that contains the term
    c. User request to see templates that contains the term
6. Optional research return results:
    a. System returns ontology viewer and jumps to the term
    b. System returns document viewer showing all related documents, and highlight the term
    c. returns template viewer showing all related templates
7. User may continue other optional search in returned viewers

The above shows a typical use case to represent user authentication (step 1 and 2), a single term search (step 3), retrieval of results (step 4) and optional further search (step 5 and 6).

Multi-condition query

1. Same login process as above
2. User use query builder to submit a complex query, e.g. if the user wants to search for radiology report(s) that have *complete tear at posterior horn of lateral meniscus* or *intact ACL*:
    a. Set general group condition relation: **OR**
    b. Add sub condition group and set sub group internal relation: **AND**
    c. Choose search field: **Anatomy**, and fill keyword field: ***lateral meniscus***
    d. Choose search filed: **Anatomy qualifier**, and fill keyword field: ***posterior horn***
    e. Choose search field: **Finding**, and fill keyword field: ***tear***
    f. Choose search field: **Finding qualifier**, and fill keyword field: ***complete***
    g. Add a parallel group to the previous one, and set group internal relation: AND
    h. Choose search field: **Finding**, and fill keyword field: ***intact***
    i. Choose search field: **Anatomy**, and fill keyword field: ***ACL***
3. System return template viewer showing related templates
4. Optional further research from the template viewer

The above shows a typical use case of submit a multi-condition query and retrieval of results. Search fields are predefined as mentioned in ***Section 7.2.2***. It is not necessarily to choose a specific search field if the user is unsure about the category of intended keyword. The search field **Term** can be chosen instead. Then the system will search for any slot that matches the given keyword.

Chapter 8 Conclusion

The aim of research presented in this thesis is to automate semantic interpretation on clinical narratives. The intention is that by overcoming obstacles caused by costly manual efforts with automated interpretation processes on clinical narratives, we will be able to reuse clinical narratives to support epidemiologic research.

The use of specific document types (knee MRI reports) and ontology (TRAK) allowed us to test a hypothesis about the feasibility of our approach to rapid development of high performing IE systems. Nonetheless, the approach itself is not restricted to a particular domain and as such represents a general contribution to computer science in addition to project deliverables that include an extended ontology and two systems it enables to perform information extraction (KneeTex) and information retrieval (KneeBase).

Our achievements include: a reusable rapid ontology development framework demonstrated by expanding TRAK; an expanded TRAK ontology; an ontology-based information extraction system KneeTex; and a web-based information retrieval system KneeBase. ***Section 8.1*** reviews achievements in this research, followed by a discussion of limitations in ***Section 8.2***. Possible future work is discussed in ***Section 8.3***.

8.1 Summary of contributions

8.1.1 **Rapid ontology development framework**

We adopted an alternative approach based on a set of strategies that can be used to systematically expand the coverage of existing ontologies or to develop them from scratch. Three of these strategies are data–driven and as such are more likely to ensure that the ontology effectively supports the intended NLP application. Each data–driven strategy utilises a different approach to extracting the relevant terminology from the data either manually or automatically. The fourth strategy is based on integration of concepts from other relevant knowledge sources. The two main aims of this strategy are: (1) to avoid the overfitting of the ontology to limited data available, and (2) to provide an initial taxonomic structure to incorporate new concepts.

In this study, we illustrated how these strategies were implemented in practice to expand the coverage of the TRAK ontology to make it suitable for a specific NLP application.

8.1.2 **Expanded TRAK ontology**

We practically demonstrated the approach to update the TRAK ontology in order to allow interpretation of information contained in knee MRI reports. We expanded TRAK into a fine-grained lexicalised ontology. The expanded TRAK contains 1,621 concepts, 2,550 synonyms and 560 relationship instances, compared with 1,292 concepts, 1,720 synonyms and 518 relationships in the original one.

The TRAK ontology-based information extraction system KneeTex exhibited human-level performance, which proves that the TRAK ontology has adequate coverage and been highly attuned to the given sublanguage.

8.1.3 **KneeTex**

We developed KneeTex as an ontology–driven system for information extraction from narrative reports that describe an MRI scan of the knee. The system exhibited human–level performance on a gold standard, attributed partly to the use of expanded TRAK ontology, which serves as a very fine-grained lexico-semantic knowledge base and plays a pivotal role in guiding and constraining text analysis.

The evaluation results confirm that KneeTex succeeded in making effective use of the ontology to support information extraction from knee MRI reports.

8.1.4 **KneeBase**

We demonstrated KneeBase as an example of integrated use of ontology and extracted information to build an information retrieval system. The system structure and programming framework can be reused for other similar ontology-based information retrieval tasks. As it has been proved with many previous studies that information retrieval systems have positive impact on clinical practices, we would hope that KneeBase can be used to to support large–scale multi–faceted epidemiologic studies of knee conditions.

8.2 Generalisability and limitation

Most efforts we have put in to this research are generalisable and can be exported to be applied to other similar tasks. Here we provide a step-by-step discussion on generalisability and limitations of our work.

The rapid ontology development framework is a partially automated process. It utilises primarily data-driven strategies. The two strategies, dictionary-based and automatic term

recognition, are automated processes using off-the-shelf software tools, and therefore are highly exportable to other similar tasks and require little human operation. Human effort is inevitable as required by manual data annotation and manual terminology search, though domain expert are not necessaries required in these procedures. However, to achieve satisfactory coverage for such a specific domain, domain-specific knowledge is still essential. As mentioned, we have referenced to our domain expert for many suggestions, helping solving non-agreed annotations and manual curations on suggested terms from these strategies. We used manual annotation for the purpose of identifying non-standard and infrequent terms. The reason underneath is that we have a rather small sized unannotated training set. A large and proper annotated training set would have provided sufficient information in the first two automated strategies.

The KneeTex information extraction system structure can be reused or referenced for similar tasks. However, we have used a large amount of lexical patterns and rules for term extraction and ambiguity resolution. These patterns and rules are propably restricted to the domain of knee injury MRI reports. Although we have not evaluated such patterns and rules on other dataset, it is not recommended to apply KneeTex directly on other domains without prior domain adaption.

8.3 Future work

Although our system achieved satisfactory performance, it also raises some requirements of future work, which could still be done to solve errors and further improve performance.

Domain adapted POS tagger

As mentioned in Section 6.3, POS tagging was not included in our linguistic pre-processes. This is due to declined performances of general POS taggers when applied on clinical domains. Correct assigned POS tags could be helpful to solve errors as discussed in ***Section 6.9.2***. Ferraro et al have seen an increase of 6.2% to 11.4% in accuracy after performed domain adaption processes (Ferraro et al., 2013). Therefore, more effort can be put into this.

Dependency based negation resolution

Negation resolution now assumes that findings that locate to the right of the negation term and are within the same segment with the negation term are negated. Although it worked quite well in our system, it would be better to have dependency relations involved for more accurate identification of negated concepts.

Flexible co-reference resolution

Our current system uses a simple heuristic semantic-rule-based approach to solve co-reference problems. Exceptions have caused a few unsolved co-references in our system as our rules are not flexible enough. Many hybrid and supervised co-reference resolution systems participated and performed well in the 2011 i2b2 co-reference challenges (Uzuner et al., 2012). To introduce supervised machine learning to create flexible co-reference resolution is worth considering for future work.

Reduce human effort

Human effort still takes a significant part in this project. Manual annotation, terminology search and curation all require some human interferences and domain knowledge. Although at current stage human effort is to some extent required, it is still possible to reduce some human effort with the availability of more data sets, which would improve the performance of the two automated strategies we have. There also exist many rule induction algorithms, which could potentially be applied to replace part of human effort in recognising rules and patterns.

8.4 Summary

In this research, we successfully tested our hypotheses that semantic interpretation of clinical narratives can be automated with ontology-based text mining and it is also feasible to expand an existing ontology effectively in a systematic pipeline.

We achieved satisfactory performance in the strategies we developed to expand the TRAK ontology. Based on this expanded TRAK ontology, we successfully automated the semantic interpretation process with support from ontology-based text mining.

Most of outcomes in this research can be reused or referenced in future research, including a rapid ontology development framework, the structure of our information extraction system KneeTex, the expanded TRAK ontology and a reusable information retrieval system framework.

Although there is still room for improvement, our research demonstrated the state-of-the-art performance. We hope that this research could provide useful insights for future text mining support on epidemiologic studies.

Bibliography

[1]. Abbe, A., Grouin, C., Zweigenbaum, P., Falissard, B., 2015. Text mining applications in psychiatry: a systematic literature review. Int. J. Methods Psychiatr. Res. 25, 86–100. doi:10.1002/mpr.1481

[2]. Adamusiak, T., Burdett, T., Kurbatova, N., Joeri van der Velde, K., Abeygunawardena, N., Antonakaki, D., Kapushesky, M., Parkinson, H., Swertz, M.A., 2011. OntoCAT--simple ontology search and integration in Java, R and REST/JavaScript. BMC Bioinformatics 12, 218. doi:10.1186/1471-2105-12-218

[3]. Al-Safadi, L., Alomran, R., Almutairi, F., 2013. Evaluation of Metamap Performance in Radiographic Images Retrieval. Research Journal of Applied Sciences, Engineering and Technology 22, 4231–4236.

[4]. Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W.F., Warner, C., Hwang, J.D., Choi, J.D., Dligach, D., Nielsen, R.D., Martin, J., Ward, W., Palmer, M., Savova, G.K., 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Proc AMIA Annu Fall Symp 20, 922–930. doi:10.1136/amiajnl-2012-001317

[5]. Alias-i, 2008. LingPipe 4.1.0 [Online] Alias-i. Available at: http://alias-i.com/lingpipe (accessed 12.13).

[6]. Aronson, A.R., 2006. MetaMap: Mapping Text to the UMLS Metathesaurus 1–26.

[7]. Aronson, A.R., 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 17–21.

[8]. Aronson, A.R., 1996. The effect of textual variation on concept based information retrieval. Proc AMIA Annu Fall Symp 373–377.

[9]. Aronson, A.R., Lang, F.-M., 2010. An overview of MetaMap: historical perspective and recent advances. 17, 229–236. doi:10.1136/jamia.2009.002733

[10]. Arthritis Research UK, 2013. Osteoarthritis in General Practice [Online] Available at: http://www.arthritisresearchuk.org/arthritis-information/data-and-statistics/osteoarthritis.aspx (accessed 5.20.16).

[11]. Artstein, R., Poesio, M., 2008. Inter-Coder Agreement for Computational Linguistics. Computational Linguistics 34, 555–596.

[12]. Barrows, R.C., Jr, Busuioc, M., Friedman, C., 2000. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proc AMIA Symp 51–55.

[13]. Bellamy, N., Buchanan, W.W., Goldsmith, C.H., Campbell, J., Stitt, L.W., 1988. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J. Rheumatol. 15, 1833–1840.

[14]. Bernard, E., Arnould, M., Saint-Lary, O., Duhot, D., Hebbrecht, G., 2012. Internet use for information seeking in clinical practice: A cross-sectional survey among French general practitioners. IJMI 81, 493–499. doi:10.1016/j.ijmedinf.2012.02.001

[15]. Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 284, 34–.

[16]. Bodenreider, O., 2004. The Unified Medical Language System (UMLS):

integrating biomedical terminology. Nucleic Acids Res. 32, D267–70. doi:10.1093/nar/gkh061

[17]. Bourhis, R.Y., Roth, S., MacQueen, G., 1989. Communication in the hospital setting: a survey of medical and everyday language use amongst patients, nurses and doctors. Soc Sci Med 28, 339–346.

[18]. Bullinaria, J.A., 2005. Semantic Networks and Frames 1–20.

[19]. Buntin, M.B., Burke, M.F., Hoaglin, M.C., Blumenthal, D., 2011. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Aff (Millwood) 30, 464–471. doi:10.1377/hlthaff.2011.0178

[20]. Button, K., Iqbal, A.S., Letchford, R.H., van Deursen, R.W.M., 2012. Clinical effectiveness of knee rehabilitation techniques and implications for a self-care treatment model. Physiotherapy 98, 288–299. doi:10.1016/j.physio.2011.08.003

[21]. Button, K., Roos, P.E., van Deursen, R.W.M., 2014. Activity progression for anterior cruciate ligament injured individuals. Clin Biomech (Bristol, Avon) 29, 206–212. doi:10.1016/j.clinbiomech.2013.11.010

[22]. Button, K., van Deursen, R.W., Soldatova, L., Spasić, I., 2013. TRAK ontology: Defining standard care for the rehabilitation of knee conditions. Journal of Biomedical Informatics 46, 615–625. doi:10.1016/j.jbi.2013.04.009

[23]. Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376. doi:10.1038/nrn3475

[24]. Cambria, E., White, B., 2014. Jumping NLP Curves: A Review of Natural Language Processing Research. Ieee Computational Intelligence Magazine 9, 48–57. doi:10.1109/MCI.2014.2307227

[25]. Campbell, D.A., Johnson, S.B., 2001. Comparing syntactic complexity in medical and non-medical corpora. Proc AMIA Symp 90–94.

[26]. Can’t Beat Jazzy: Introducing the Java Platform’s Jazzy New Spell Checker API, 2013. Can’t Beat Jazzy: Introducing the Java Platform’s Jazzy New Spell Checker API.

[27]. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G., 2001a. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 34, 301–310. doi:10.1006/jbin.2001.1029

[28]. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G., 2001b. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Annu Fall Symp 105–109.

[29]. Chapman, W.W., Nadkarni, P.M., Hirschman, L., D'Avolio, L.W., Savova, G.K., Uzuner, Ö., 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 18, 540–543. doi:10.1136/amiajnl-2011-000465

[30]. Chen, Y.S., Chong, P.P., Tong, M.Y., 1994. Mathematical and computer modelling of the Pareto principle. Mathematical and Computer Modelling: An International Journal 19, 61–80. doi:10.1016/0895-7177(94)90041-8

[31]. Christensen, L.M., Haug, P.J., Fiszman, M., 2002. MPLUS: a probabilistic medical

language understanding system, in:. Presented at the BioMed '02: Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain, Association for Computational Linguistics.

[32]. Ciravegna, F., 1995. Understanding messages in a diagnostic domain. Information Processing & Management 31, 687–701. doi:10.1016/0306-4573(95)00027-E

[33]. Clark, C., Good, K., Jezierny, L., Macpherson, M., Wilson, B., Chajewska, U., 2008. Identifying smokers with a medical extraction system. Proc AMIA Annu Fall Symp 15, 36–39. doi:10.1197/jamia.M2442

[34]. Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power-Law Distributions in Empirical Data. SIAM Review 51, 661–703. doi:10.1137/070710111

[35]. Clayton, R.A.E., Court-Brown, C.M., 2008. The epidemiology of musculoskeletal tendinous and ligamentous injuries. Injury 39, 1338–1344. doi:10.1016/j.injury.2008.06.021

[36]. Cohen, W., Ravikumar, P., Fienberg, S., 2003. A Comparison of String Distance Metrics for Name-Matching Tasks.

[37]. Cohen, W.W., 2004. *MinorThird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data*.

[38]. Collins, A.M., Quillian, M.R., 1969. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240–247.

[39]. Cowie, J., Lehnert, W., 1996. Information extraction. Communications of the ACM 39, 80–91. doi:10.1145/234173.234209

[40]. Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H., 2006. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics 7, 97. doi:10.1186/1471-2105-7-97

[41]. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P., 1992. A practical part-of-speech tagger, in:. Presented at the the third conference, Association for Computational Linguistics, Morristown, NJ, USA, pp. 133–140. doi:10.3115/974499.974523

[42]. Damerau, F.J., 1964. A Technique for Computer Detection and Correction of Spelling Errors. Communications of the ACM 7, 171–176. doi:10.1145/363958.363994

[43]. Day-Richter, J., Harris, M.A., Haendel, M., Group, T.G.O.O.-E.W., Lewis, S., 2007. OBO-Edit—an ontology editor for biologists. Bioinformatics 23, 2198–2200. doi:10.1093/bioinformatics/btm112

[44]. EMBL-EBI, 2016. Ontology Lookup Service [Online] EMBL-EBI. Available at: http://www.ebi.ac.uk/ols (accessed 16).

[45]. Fan, J.-W., Prasad, R., Yabut, R.M., Loomis, R.M., Zisook, D.S., Mattison, J.E., Huang, Y., 2011. Part-of-speech tagging for clinical text: wall or bridge between institutions? AMIA Annu Symp Proc 2011, 382–391. doi:10.1016/j.jbi.2011.04.006

[46]. Fan, J.-W., Yang, E.W., Jiang, M., Prasad, R., Loomis, R.M., Zisook, D.S., Denny, J.C., Xu, H., Huang, Y., 2013. Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. J Am Med Inform Assoc 20, 1168–1177. doi:10.1136/amiajnl-2013-001810

[47]. Federhen, S., 2012. The NCBI Taxonomy database. Nucleic Acids Res. 40, D136–43. doi:10.1093/nar/gkr1178

[48]. Fernandez-Lopez, M., Gomez-Perez, A., 2002. Overview and analysis of methodologies for building ontologies. Knowledge Engineering Review 17, 129–156. doi:10.1017/S0269888902000462

[49]. Ferraro, J.P., Daumé, H., III, DuVall, S.L., Chapman, W.W., Harkema, H., Haug, P.J., 2013. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. J Am Med Inform Assoc 20, 931–939. doi:10.1136/amiajnl-2012-001453

[50]. Ferrucci, D., Lally, A., 2004. Building an example application with the Unstructured Information Management Architecture. IBM Systems Journal 43, 455–475. doi:10.1147/sj.433.0455

[51]. Fiszman, M., Blatter, D.D., Christensen, L.M., Oderich, G., Macedo, T., Eidelwein, A.P., Haug, P.J., 2002. Utilization review of head CT scans: value of a medical language processing system.

[52]. Fleiss, J.L., 1971. Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin 76, 378–382.

[53]. Friedlin, J., Overhage, M., 2011. An evaluation of the UMLS in representing corpus derived clinical concepts. Proc AMIA Symp 2011, 435–444.

[54]. Friedman, C., 2006. Semantic Text Parsing for Patient Records, in: Chen, H., Fuller, S.S., Friedman, C., Hersh, W. (Eds.), Medical Informatics. Springer US, pp. 423–448. doi:10.1007/0-387-25739-X_15

[55]. Friedman, C., 2000. A broad-coverage natural language processing system. Proc AMIA Symp 270–274.

[56]. Friedman, C., 1992. The UMLS coverage of clinical radiology. Proceedings of the Annual Symposium on Computer Application in Medical Care 309.

[57]. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., Johnson, S.B., 1994. A general natural-language text processor for clinical radiology. 1, 161–174.

[58]. Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S.B., Clayton, P.D., 2008. Natural language processing in an operational clinical information system. Nat. Lang. Eng. 1, 1–27. doi:10.1017/S1351324900000061

[59]. Friedman, C., Kra, P., Rzhetsky, A., 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics 35, 222–235. doi:10.1016/S1532-0464(03)00012-1

[60]. Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K.B., Hunter, L.E., Verspoor, K., 2014. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinformatics 15, 59. doi:10.1186/1471-2105-15-59

[61]. Gage, B.E., McIlvain, N.M., Collins, C.L., Fields, S.K., Dawn Comstock, R., 2012. Epidemiology of 6.6 Million Knee Injuries Presenting to United States Emergency Departments From 1999 Through 2008. Academic Emergency Medicine 19, 378–385. doi:10.1111/j.1553-2712.2012.01315.x

[62]. Geertzen, J., 2012. Inter-Rater Agreement with multiple raters and variables [Online] URL https://mlnl.net/jg/software/ira/ (accessed 15).

[63]. Gow, J., 2009. Artificial Intelligence.

[64]. Grenon, P., 2008. Chapter 3: A Primer on Knowledge Representation and Ontological Engineering, in: Applied Ontology, An Introduction. DE GRUYTER,

Berlin, Boston, pp. 57–81. doi:10.1515/9783110324860.57

[65]. Grenon, P., Smith, B., Goldberg, L., 2004. Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform 102, 20–38.

[66]. Grishman, R., 1997. Information Extraction: Techniques and Challenges, in: Pazienza, M.T. (Ed.), Information Extraction. Springer, Heidelberg, pp. 1–18.

[67]. Grishman, R., Sundheim, B., 1996. Message Understanding Conference - 6: A Brief History, in:. Presented at the International Conference on Computational Linguistics, pp. 466–471.

[68]. Grover, M., 2012. Evaluating acutely injured patients for internal derangement of the knee. Am Fam Physician 85, 247–252.

[69]. Guermazi, A., Niu, J., Hayashi, D., Roemer, F.W., Englund, M., Neogi, T., Aliabadi, P., McLennan, C.E., Felson, D.T., 2012. Prevalence of abnormalities in knees detected by MRI in adults without knee osteoarthritis: population based observational study (Framingham Osteoarthritis Study). BMJ 345, e5339–e5339. doi:10.1136/bmj.e5339

[70]. Hammond, L.E., Lilley, J., Ribbans, W.J., 2009. Coding sports injury surveillance data: has version 10 of the Orchard Sports Injury Classification System improved the classification of sports medicine diagnoses? Br J Sports Med 43, 498–502. doi:10.1136/bjsm.2008.051979

[71]. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W., 2009. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics 42, 839–851. doi:10.1016/j.jbi.2009.05.002

[72]. Harris, Z.S., 1991. A theory of language and information. Oxford University Press, USA.

[73]. Hastie, T., Tibshirani, R., Friedman, J., 2008. Unsupervised Learning, in: The Elements of Statistical Learning, Springer Series in Statistics. Springer New York, New York, NY, pp. 1–101. doi:10.1007/b94608_14

[74]. Hayes, P.J., 1983. The Second Naive Physics Manifesto.

[75]. Herre, H., 2010. General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling, in: Theory and Applications of Ontology: Computer Applications. Springer Netherlands, pp. 297–345. doi:10.1007/978-90-481-8847-5_14

[76]. Hersh, W.R., Campbell, E.M., Malveau, S.E., 1997. Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: A lexical analysis. Proc AMIA Annu Fall Symp 580–584.

[77]. Hirschman, L., Sager, N., 2002. Chapter 2. Automatic Information Formatting of a Medical Sublanguage, in: Kittredge, R., Lehrberger, J. (Eds.), Sublanguage, Studies of Language in Restricted Semantic Domains. DE GRUYTER, Berlin, Boston. doi:10.1515/9783110844818-003

[78]. Homayouni, R., Heinrich, K., Wei, L., Berry, M.W., 2005. Gene clustering by Latent Semantic Indexing of MEDLINE abstracts. Bioinformatics 21, 104–115. doi:10.1093/bioinformatics/bth464

[79]. Hong, Y., Zhang, J., Heilbrun, M.E., Kahn, C.E., Jr, 2012. Analysis of RadLex Coverage and Term Co-occurrence in Radiology Reporting Templates. J Digit

Imaging 25, 56–62. doi:10.1007/s10278-011-9423-9

[80]. Horridge, M., Bechhofer, S., 2011. The OWL API: A Java API for OWL ontologies. Semantic Web 2, 11–21. doi:10.3233/SW-2011-0025

[81]. Hripcsak, G., Austin, J.H.M., Alderson, P.O., Friedman, C., 2002. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 224, 157–163. doi:10.1148/radiol.2241011118

[82]. Hripcsak, G., Rothschild, A.S., 2005. Agreement, the f-measure, and reliability in information retrieval. 12, 296–298. doi:10.1197/jamia.M1733

[83]. Hsiao, C.-J., Hing, E., Socey, T.C., Cai, B., 2010. Electronic Medical Record/Electronic Health Record Systems of Office-based Physicians: United States, 2009 and Preliminary 2010 State Estimates 1–6.

[84]. IEEE, 1998. IEEE Standard for Developing Software Life Cycle Processes. IEEE, Piscataway, NJ, USA. doi:10.1109/IEEESTD.1998.88827

[85]. ihtsdo, 2014. SNOMED CT Starter Guide.

[86]. Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2, e124. doi:10.1371/journal.pmed.0020124

[87]. Jacobson, 1999. The Unified Software Development Process. Pearson Education India.

[88]. Javaid, M.K., Lynch, J.A., Tolstykh, I., Guermazi, A., Roemer, F., Aliabadi, P., McCulloch, C., Curtis, J., Felson, D., Lane, N.E., Torner, J., Nevitt, M., 2010. Pre-radiographic MRI findings are associated with onset of knee symptoms: the most study. Osteoarthr. Cartil. 18, 323–328. doi:10.1016/j.joca.2009.11.002

[89]. Jensen, P.B., Jensen, L.J., Brunak, S., 2012. Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. 13, 395–405. doi:10.1038/nrg3208

[90]. Jiang, J., 2008. Domain Adaptation in Natural Language Processing. ProQuest.

[91]. Jiang, M., Huang, Y., Fan, J.-W., Tang, B., Denny, J., Xu, H., 2015. Parsing clinical text: how good are the state-of-the-art parsers? MIDM) 13(S-1 15, 1. doi:10.1186/1472-6947-15-S1-S2

[92]. Jimeno-Yepes, A., Sticco, J.C., Mork, J.G., Aronson, A.R., 2013. GeneRIF indexing: sentence selection based on machine learning. BMC Bioinformatics 14, 171. doi:10.1186/1471-2105-14-171

[93]. Jonnalagadda, S., Cohen, T., Wu, S.T.-I., Gonzalez, G., 2012. Enhancing clinical concept extraction with distributional semantics. Journal of Biomedical Informatics 45, 129–140. doi:10.1016/j.jbi.2011.10.007

[94]. Justeson, J.S., Katz, S.M., 2008. Technical terminology: some linguistic properties and an algorithm for identification in text. Nat. Lang. Eng. 1. doi:10.1017/S1351324900000048

[95]. Kalibatiene, D., Vasilecas, O., 2011. Survey on Ontology Languages, in: Grabis, J., Kirikova, M. (Eds.), Perspectives in Business Informatics Research, Lecture Notes in Business Information Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 124–141. doi:10.1007/978-3-642-24511-4_10

[96]. Konan, S., Rayan, F., Haddad, F.S., 2009. Do physical diagnostic tests accurately detect meniscal tears? Knee Surg Sports Traumatol Arthrosc 17, 806–811.

doi:10.1007/s00167-009-0803-3

[97]. Krovetz, R., 1997. Homonymy and polysemy in information retrieval, in:. Presented at the EACL '97: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp. 72–79. doi:10.3115/979617.979627

[98]. Kwong, R.Y., Yucel, E.K., 2003. Computed tomography scan and magnetic resonance imaging, Circulation. doi:10.1161/01.CIR.0000086899.32832.EC

[99]. L Sumathy, K., Chidambaram, M., 2013. Text Mining: Concepts, Applications, Tools and Issues An Overview. IJCA 80, 29–32. doi:10.5120/13851-1685

[100]. Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

[101]. Langlotz, C.P., 2006. RadLex: a new method for indexing online educational materials. Radiographics 26, 1595–1597. doi:10.1148/rg.266065168

[102]. Lehrberger, J., 2014. Sublanguage Analysis, in: Analyzing Language in Restricted Domains. Psychology Press, pp. 19–38.

[103]. Leroy, G., Chen, H., 2001. Meeting medical terminology needs-the ontology-enhanced Medical Concept Mapper. TITB 5, 261–270. doi:10.1109/4233.966101

[104]. Lewis, P., 2010. Current Trends in Information Technology.

[105]. Lipscomb, C.E., 2000. Medical Subject Headings (MeSH). Bulletin of the Medical Library Association 88, 265–266.

[106]. Lison, P., 2012. An introduction to machine learning 1–35.

[107]. Liu, L., Özsu, M.T. (Eds.), 2009. Encyclopedia of Database Systems. Springer US, Boston, MA. doi:10.1007/978-0-387-39940-9

[108]. Luyten, F.P., Denti, M., Filardo, G., Kon, E., Engebretsen, L., 2012. Definition and classification of early osteoarthritis of the knee. Knee Surg Sports Traumatol Arthrosc 20, 401–406. doi:10.1007/s00167-011-1743-2

[109]. Maedche, A., 2012. Ontology Learning for the Semantic Web. Springer Science & Business Media, Boston, MA. doi:10.1007/978-1-4615-0925-7

[110]. Maimon, O., Rokach, L., 2005. Introduction to Supervised Methods, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, New York, pp. 149–164. doi:10.1007/0-387-25465-X_8

[111]. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D., 2014. The Stanford CoreNLP Natural Language Processing Toolkit, in:. Presented at the 52nd Annual Meeting of the Association for Computational Linguistics System Demonstrations, pp. 55–60.

[112]. Mansouri, A., Affendey, L.S., Mamat, A., 2008. Named entity recognition approaches. International Journal of Computer Science and Network Security.

[113]. Manzoor, U., Usman, M., A, M., Mueen, A., 2015. Ontology-Based Clinical Decision Support System for Predicting High-Risk Pregnant Woman. International Journal of Advanced Computer Science and Applications 6. doi:10.14569/IJACSA.2015.061228

[114]. Mate, S., Koepcke, F., Toddenroth, D., Martin, M., Prokosch, H.-U., Buerkle, T., Ganslandt, T., 2015. Ontology-Based Data Integration between Clinical and

Research Systems. PLoS ONE 10. doi:10.1371/journal.pone.0116656

[115]. Maxwell, J.L., Keysor, J.J., Niu, J., Singh, J.A., Wise, B.L., Frey-Law, L., Nevitt, M.C., Felson, D.T., 2013. Participation following knee replacement: the MOST cohort study. Phys Ther 93, 1467–1474. doi:10.2522/ptj.20130109

[116]. Maynard, D., Peters, W., Li, Y., 2006. Metrics for Evaluation of Ontology-based Information Extraction, in:. Presented at the WWW Workshop on Evaluation of Ontologies for the Web, pp. 1–8.

[117]. McCarthy, J., 1987. Generality in artificial intelligence. Communications of the ACM 30, 1030–1035. doi:10.1145/33447.33448

[118]. McCray, A.T., Srinivasan, S., Browne, A.C., 1994. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 235–239.

[119]. Merriam-Webster, n.d. Merriam-Webster Medical Dictionary [Online] Merriam-Webster. URL www.merriam-webster.com (accessed 12).

[120]. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 128–144.

[121]. Moens, M.-F., 2006. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer Science & Business Media. doi:10.1007/978-1-4020-4993-4

[122]. Mohanty, S.K., Piccoli, A.L., Devine, L.J., Patel, A.A., William, G.C., Winters, S.B., Becich, M.J., Parwani, A.V., 2007. Synoptic tool for reporting of hematological and lymphoid neoplasms based on World Health Organization classification and College of American Pathologists checklist. BMC Cancer 7, 144. doi:10.1186/1471-2407-7-144

[123]. mongoDB, 2015a. mongoDB [Online] Available at: http://www.mongodb.org/ (accessed 15a).

[124]. mongoDB, 2015b. MongoDB CRUD Introduction [Online] MongoDB manual. Available at: http://docs.mongodb.org/manual/core/crud-introduction/ (accessed 9.1.15b).

[125]. Muslea, I., 1999. Extraction patterns for information extraction tasks: A survey. The AAAI-99 Workshop on Machine Learning for Information Extraction.

[126]. Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes 30, 3–26. doi:10.1075/li.30.1.03nad

[127]. National Health Service 2015. Martha Lane Fox sets out key digital proposals for the NHS [Online] Available at: https://www.england.nhs.uk/2015/12/martha-lane-fox/ [Accessed: December 2015].

[128]. NCBO, 2015. NCBO Widgets [Online] NCBO Public WIKI. Available at: http://www.bioontology.org/wiki/index.php/NCBO_Widgets (accessed 15).

[129]. NCBO, 2013. BioPortal [Online] Available at: http://bioportal.bioontology.org/ (accessed 15).

[130]. Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.R., 1991. Enabling Technology for Knowledge Sharing. Ai Magazine 12, 36–56.

[131]. Nkwenti-Azeh, B., 2001. Terminology, in: Chan, S.-W., Pollard, D.E. (Eds.), An Encyclopedia of Translation. Chiniese University Press, pp. 610–627.

[132]. NUANCE, 2008. The Electronic Patient Narrative 1–12.

[133]. OpenNLP, 2010. Apache OpenNLP [Online] The Apache Software Foundation. URL https://opennlp.apache.org (accessed 3.12).

[134]. Patterson, O., Hurdle, J.F., 2011. Document clustering of clinical narratives: a systematic study of clinical sublanguages. AMIA Annu Symp Proc 2011, 1099–1107.

[135]. Pessis, E., Drapé, J.-L., Ravaud, P., Chevrot, A., Dougados, M., Ayral, X., 2003. Assessment of progression in knee osteoarthritis: results of a 1 year study comparing arthroscopy and MRI. Osteoarthr. Cartil. 11, 361–369.

[136]. Philips, L., 1990. Hanging on the Metaphone. Computer Language magazine 7, 39–44.

[137]. Piskorski, J., Yangarber, R., 2013. Information Extraction: Past, Present and Future, in: Multi-Source, Multilingual Information Extraction and Summarization. pp. 23–49. doi:10.1007/978-3-642-28569-1__2

[138]. Pompan, D.C., 2012. Reassessing the role of MRI in the evaluation of knee pain. Am Fam Physician 85, 221–224.

[139]. Poole, D.L., Mackworth, A.K., 2010. Artificial Intelligence. Cambridge University Press, Cambridge. doi:10.1017/CBO9780511794797

[140]. Radiological Society of North America, 2012. RSNA Radiology Reporting Templates [Online] RSNA Informatics Reporting. Available at: http://www.radreport.org/template/0000057 (accessed 8.17.15).

[141]. Rae, K., Orchard, J., 2007. The Orchard Sports Injury Classification System (OSICS) version 10. Clin J Sport Med 17, 201–204. doi:10.1097/JSM.0b013e318059b536

[142]. Rahimi, A., Liaw, S.-T., Taggart, J., Ray, P., Yu, H., 2014. Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in Electronic Health Records. IJMI 83, 768–778. doi:10.1016/j.ijmedinf.2014.06.002

[143]. Raja, K., Jonnalagadda, S.R., 2015. Natural Language Processing and Data Mining for Clinical Text, in: Healthcare Data Analytics. CRC Press.

[144]. Rashbass, J., 1995. Online Mendelian Inheritance in Man. Trends Genet. 11, 291–291.

[145]. Roberts, K., Shooshan, S.E., Rodriguez, L., Abhyankar, S., Kilicoglu, H., Demner-Fushman, D., 2015. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. Journal of Biomedical Informatics 58 Suppl, S111–9. doi:10.1016/j.jbi.2015.06.010

[146]. Robinson, P.N., Koehler, S., Bauer, S., Seelow, D., Horn, D., Mundlos, S., 2008. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. Am. J. Hum. Genet. 83, 610–615. doi:10.1016/j.ajhg.2008.09.017

[147]. Roemer, F.W., Guermazi, A., Felson, D.T., Niu, J., Nevitt, M.C., Crema, M.D., Lynch, J.A., Lewis, C.E., Torner, J., Zhang, Y., 2011. Presence of MRI-detected joint effusion and synovitis increases the risk of cartilage loss in knees without osteoarthritis at 30-month follow-up: the MOST study. Ann. Rheum. Dis. 70, 1804–1809. doi:10.1136/ard.2011.150243

[148]. Roos, E.M., Roos, H.P., Lohmander, L.S., Ekdahl, C., Beynnon, B.D., 1998. Knee

Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. J Orthop Sports Phys Ther 28, 88–96. doi:10.2519/jospt.1998.28.2.88

[149]. Rosse, C., Mejino, J.L.V., 2003. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of Biomedical Informatics 36, 478–500. doi:10.1016/j.jbi.2003.11.007

[150]. Roth, D., Zelenko, D., 1998. Part of Speech Tagging Using a Network of Linear Separators. COLING-ACL 1136–1142.

[151]. Royal College of Radiologists, 2006. Standards for the Reporting and Interpretation of Imaging Investigations.

[152]. Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J., 1994. Natural language processing and the representation of clinical data. 1, 142–160.

[153]. Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management 24, 513–523. doi:10.1016/0306-4573(88)90021-0

[154]. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G., 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. 17, 507–513. doi:10.1136/jamia.2009.001560

[155]. Schattner, P., Saunders, M., Stanger, L., Speak, M., Russo, K., 2010. Clinical data extraction and feedback in general practice: a case study from Australian primary care. jhi 18, 205–212. doi:10.14236/jhi.v18i3.773

[156]. Scheuermann, R.H., Ceusters, W., Smith, B., 2009. Toward an ontological treatment of disease and diagnosis. Summit on Translat Bioinforma 2009, 116–120.

[157]. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.-W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A., 2012. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 40, 940–946. doi:10.1093/nar/gkr972

[158]. Sheeba, J.I., Vivekanandan, K., Sabitha, G., Padmavathi, P., 2013. Unsupervised Hidden Topic Framework for Extracting Keywords (Synonym, Homonym, Hyponymy and Polysemy) and Topics in Meeting Transcripts, in: Advances in Computing and Information Technology. Springer Berlin Heidelberg, pp. 299–307. doi:10.1007/978-3-642-31552-7_32

[159]. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25, 1251–1255. doi:10.1038/nbt1346

[160]. Sonin, A.H., Fitzgerald, S.W., Bresler, M.E., Kirsch, M.D., Hoff, F.L., Friedman, H., 1995. MR imaging appearance of the extensor mechanism of the knee: functional anatomy and injury patterns. Radiographics 15, 367–382. doi:10.1148/radiographics.15.2.7761641

[161]. Sorel, D., 2015. jQuery QueryBuilder.

[162]. Spasic, I., Ananiadou, S., McNaught, J., Kumar, A., 2005. Text mining and ontologies in biomedicine: Making sense of raw text. Brief. Bioinformatics 6, 239–

251.

[163]. Spasić, I., Greenwood, M., Preece, A., Francis, N., Elwyn, G., 2013. FlexiTerm: a flexible term recognition method. J Biomed Semantics 4, 27. doi:10.1186/2041-1480-4-27

[164]. Spasić, I., Livsey, J., Keane, J.A., Nenadić, G., 2014. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 83, 605–623. doi:10.1016/j.ijmedinf.2014.06.009

[165]. Stanford Centre for Biomedical Informatics Research, n.d. Protégé.

[166]. Starren, J., Johnson, S.M., 1996. Notations for high efficiency data presentation in mammography. Proceedings of the AMIA Annual Fall Symposium 557.

[167]. Stede, M., 2000. The hyperonym problem revisited: conceptual and lexical hierarchies in language generation, in:. Presented at the Proceedings of the first international conference on Natural language generation, Association for Computational Linguistics, Morristown, NJ, USA, pp. 93–99. doi:10.3115/1118253.1118267

[168]. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J., 2012. BRAT: a web-based tool for NLP-assisted text annotation, in:. Presented at the EACL '12: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics.

[169]. Stetson, P.D., Johnson, S.B., Scotch, M., Hripcsak, G., 2002. The sublanguage of cross-coverage. Proceedings of the AMIA Symposium 742.

[170]. Stubbs, A., Kotfila, C., Xu, H., Uzuner, Ö., 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. Journal of Biomedical Informatics 58 Suppl, S67–77. doi:10.1016/j.jbi.2015.07.001

[171]. Taira, R.K., 2009. Natural Language Processing of Medical Reports, in: Bui, A.A.T., Taira, R.K. (Eds.), Medical Imaging Informatics. Springer US, Boston, MA, pp. 257–298. doi:10.1007/978-1-4419-0385-3_6

[172]. Tanenblatt, M.A., Coden, A., Sominsky, I.L., 2010. The ConceptMapper Approach to Named Entity Recognition. LREC.

[173]. Tang, B., Cao, H., Wu, Y., Jiang, M., Xu, H., 2013. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. MIDM) 13(S-1 13 Suppl 1, S1. doi:10.1186/1472-6947-13-S1-S1

[174]. Tegner, Y., Lysholm, J., 1985. Rating Systems in the Evaluation of Knee Ligament Injuries. Clin. Orthop. Relat. Res. 198, 43–49.

[175]. The Gene Ontology Consortium, 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049–D1056. doi:10.1093/nar/gku1179

[176]. Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. HLT-NAACL 2003.

[177]. Toutanova, K., Manning, C.D., 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger, in:. Presented at the the 2000 Joint SIGDAT conference, Association for Computational Linguistics, Morristown, NJ, USA, pp. 63–70. doi:10.3115/1117794.1117802

[178]. Tsuruoka, Y., McNaught, J., Tsujii, J., Ananiadou, S., 2007. Learning string

similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics 23, 2768–2774. doi:10.1093/bioinformatics/btm393

[179]. U.S. National Library of Medicine, 2004. Fact SheetMEDLINE®.

[180]. UK Department of Health, 2003. Confidentiality.

[181]. UMLS, n.d. UMLS Fact Sheet [Online] U.S. National Library of Medicine. Available at: http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html (accessed 15).

[182]. UMLS, n.d. UMLS Fact Sheet [Online] U.S. National Library of Medicine. Available at: http://www.nlm.nih.gov/pubs/factsheets/umlslex.html (accessed 15).

[183]. UMLS, n.d. UMLS Fact Sheet.

[184]. UMLS, n.d. UTS Terminology Service [Online] URL https://uts.nlm.nih.gov/ (accessed 15).

[185]. Uschold, M., King, M., Moralee, S., Zorgios, Y., 1998. The Enterprise Ontology. Knowledge Engineering Review 13, 31–89. doi:10.1017/S0269888998001088

[186]. Uschold, M., King, M., University of Edinburgh. Artificial Intelligence Applications Institute, 1995. Towards a Methodology for Building Ontologies.

[187]. Uzuner, Ö., 2009. Recognizing Obesity and Comorbidities in Sparse Data. Proc AMIA Annu Fall Symp 16, 561–570. doi:10.1197/jamia.M3115

[188]. Uzuner, Ö., Bodnari, A., Shen, S., Forbush, T., Pestian, J., South, B.R., 2012. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 19, 786–791. doi:10.1136/amiajnl-2011-000784

[189]. Uzuner, Ö., Goldstein, I., Luo, Y., Kohane, I.S., 2008. Identifying Patient Smoking Status from Medical Discharge Records. Proc AMIA Annu Fall Symp 15, 14–24. doi:10.1197/jamia.M2408

[190]. Uzuner, Ö., Luo, Y., Szolovits, P., 2007. Evaluating the state-of-the-art in automatic de-identification. Proc AMIA Annu Fall Symp 14, 550–563. doi:10.1197/jamia.M2444

[191]. Uzuner, Ö., Solti, I., Cadag, E., 2010. Extracting medication information from clinical text. J Am Med Inform Assoc 17, 514–518. doi:10.1136/jamia.2010.003947

[192]. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L., 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Proc AMIA Annu Fall Symp 18, 552–556. doi:10.1136/amiajnl-2011-000203

[193]. Uzuner, Ö., Stubbs, A., 2015. Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. Journal of Biomedical Informatics 58 Suppl, S1–5. doi:10.1016/j.jbi.2015.10.007

[194]. Uzuner, Ö., Szolovits, P., Kohane, I., 2006. i2b2 Workshop on Natural Language Processing Challenges for Clinical Records, in:. Presented at the Proceedings of the Fall Symposium of the American Medical Informatics Association.

[195]. van Ginneken, A.M., de Wilde, M., van Mulligen, E.M., Stam, H.J., 1997. Can data representation and interface demands be reconciled? Approach in ORCA. Proc AMIA Annu Fall Symp 779–783.

[196]. van Middelkoop, M., van Linschoten, R., Berger, M.Y., Koes, B.W., Bierma-Zeinstra, S.M.A., 2008. Knee complaints seen in general practice: active sport

participants versus non-sport participants. BMC Musculoskelet Disord 9. doi:10.1186/1471-2474-9-36

[197]. Vapnik, V., 1982. Estimation of Dependences Based on Empirical Data. Springer-Verlag New York, Inc.

[198]. Vos, T., Flaxman, A.D., Naghavi, M., Lozano, R., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J.A., Abdalla, S., Aboyans, V., Abraham, J., Ackerman, I., Aggarwal, R., Ahn, S.Y., Ali, M.K., Alvarado, M., Anderson, H.R., Anderson, L.M., Andrews, K.G., Atkinson, C., Baddour, L.M., Bahalim, A.N., Barker-Collo, S., Barrero, L.H., Bartels, D.H., Basáñez, M.-G., Baxter, A., Bell, M.L., Benjamin, E.J., Bennett, D., Bernabé, E., Bhalla, K., Bhandari, B., Bikbov, B., Bin Abdulhak, A., Birbeck, G., Black, J.A., Blencowe, H., Blore, J.D., Blyth, F., Bolliger, I., Bonaventure, A., Boufous, S., Bourne, R., Boussinesq, M., Braithwaite, T., Brayne, C., Bridgett, L., Brooker, S., Brooks, P., Brugha, T.S., Bryan-Hancock, C., Bucello, C., Buchbinder, R., Buckle, G., Budke, C.M., Burch, M., Burney, P., Burstein, R., Calabria, B., Campbell, B., Canter, C.E., Carabin, H., Carapetis, J., Carmona, L., Cella, C., Charlson, F., Chen, H., Cheng, A.T.-A., Chou, D., Chugh, S.S., Coffeng, L.E., Colan, S.D., Colquhoun, S., Colson, K.E., Condon, J., Connor, M.D., Cooper, L.T., Corriere, M., Cortinovis, M., de Vaccaro, K.C., Couser, W., Cowie, B.C., Criqui, M.H., Cross, M., Dabhadkar, K.C., Dahiya, M., Dahodwala, N., Damsere-Derry, J., Danaei, G., Davis, A., De Leo, D., Degenhardt, L., Dellavalle, R., Delossantos, A., Denenberg, J., Derrett, S., Jarlais, Des, D.C., Dharmaratne, S.D., Dherani, M., Diaz-Torne, C., Dolk, H., Dorsey, E.R., Driscoll, T., Duber, H., Ebel, B., Edmond, K., Elbaz, A., Ali, S.E., Erskine, H., Erwin, P.J., Espindola, P., Ewoigbokhan, S.E., Farzadfar, F., Feigin, V., Felson, D.T., Ferrari, A., Ferri, C.P., Fèvre, E.M., Finucane, M.M., Flaxman, S., Flood, L., Foreman, K., Forouzanfar, M.H., Fowkes, F.G.R., Franklin, R., Fransen, M., Freeman, M.K., Gabbe, B.J., Gabriel, S.E., Gakidou, E., Ganatra, H.A., Garcia, B., Gaspari, F., Gillum, R.F., Gmel, G., Gosselin, R., Grainger, R., Groeger, J., Guillemin, F., Gunnell, D., Gupta, R., Haagsma, J., Hagan, H., Halasa, Y.A., Hall, W., Haring, D., Haro, J.M., Harrison, J.E., Havmoeller, R., Hay, R.J., Higashi, H., Hill, C., Hoen, B., Hoffman, H., Hotez, P.J., Hoy, D., Huang, J.J., Ibeanusi, S.E., Jacobsen, K.H., James, S.L., Jarvis, D., Jasrasaria, R., Jayaraman, S., Johns, N., Jonas, J.B., Karthikeyan, G., Kassebaum, N., Kawakami, N., Keren, A., Khoo, J.-P., King, C.H., Knowlton, L.M., Kobusingye, O., Koranteng, A., Krishnamurthi, R., Lalloo, R., Laslett, L.L., Lathlean, T., Leasher, J.L., Lee, Y.Y., Leigh, J., Lim, S.S., Limb, E., Lin, J.K., Lipnick, M., Lipshultz, S.E., Liu, W., Loane, M., Ohno, S.L., Lyons, R., Ma, J., Mabweijano, J., MacIntyre, M.F., Malekzadeh, R., Mallinger, L., Manivannan, S., Marcenes, W., March, L., Margolis, D.J., Marks, G.B., Marks, R., Matsumori, A., Matzopoulos, R., Mayosi, B.M., McAnulty, J.H., McDermott, M.M., McGill, N., McGrath, J., Medina-Mora, M.E., Meltzer, M., Mensah, G.A., Merriman, T.R., Meyer, A.-C., Miglioli, V., Miller, M., Miller, T.R., Mitchell, P.B., Mocumbi, A.O., Moffitt, T.E., Mokdad, A.A., Monasta, L., Montico, M., Moradi-Lakeh, M., Moran, A., Morawska, L., Mori, R., Murdoch, M.E., Mwaniki, M.K., Naidoo, K., Nair, M.N., Naldi, L., Narayan, K.M.V., Nelson, P.K., Nelson, R.G., Nevitt, M.C., Newton, C.R., Nolte, S., Norman, P., Norman, R., O'Donnell, M., O'Hanlon, S., Olives, C., Omer, S.B., Ortblad, K., Osborne, R., Ozgediz, D., Page, A., Pahari, B., Pandian, J.D., Rivero, A.P., Patten, S.B., Pearce, N., Padilla, R.P., Perez-Ruiz, F., Perico, N., Pesudovs, K., Phillips, D., Phillips, M.R., Pierce, K., Pion, S., Polanczyk, G.V., Polinder, S., Pope, C.A., Popova, S., Porrini, E., Pourmalek, F., Prince, M., Pullan, R.L., Ramaiah, K.D., Ranganathan, D., Razavi, H., Regan, M., Rehm, J.T., Rein, D.B., Remuzzi, G., Richardson, K., Rivara, F.P.,

Roberts, T., Robinson, C., De Leòn, F.R., Ronfani, L., Room, R., Rosenfeld, L.C., Rushton, L., Sacco, R.L., Saha, S., Sampson, U., Sanchez-Riera, L., Sanman, E., Schwebel, D.C., Scott, J.G., Segui-Gomez, M., Shahraz, S., Shepard, D.S., Shin, H., Shivakoti, R., Singh, D., Singh, G.M., Singh, J.A., Singleton, J., Sleet, D.A., Sliwa, K., Smith, E., Smith, J.L., Stapelberg, N.J.C., Steer, A., Steiner, T., Stolk, W.A., Stovner, L.J., Sudfeld, C., Syed, S., Tamburlini, G., Tavakkoli, M., Taylor, H.R., Taylor, J.A., Taylor, W.J., Thomas, B., Thomson, W.M., Thurston, G.D., Tleyjeh, I.M., Tonelli, M., Towbin, J.A., Truelsen, T., Tsilimbaris, M.K., Ubeda, C., Undurraga, E.A., van der Werf, M.J., van Os, J., Vavilala, M.S., Venketasubramanian, N., Wang, M., Wang, W., Watt, K., Weatherall, D.J., Weinstock, M.A., Weintraub, R., Weisskopf, M.G., Weissman, M.M., White, R.A., Whiteford, H., Wiersma, S.T., Wilkinson, J.D., Williams, H.C., Williams, S.R.M., Witt, E., Wolfe, F., Woolf, A.D., Wulf, S., Yeh, P.-H., Zaidi, A.K.M., Zheng, Z.-J., Zonies, D., Lopez, A.D., Murray, C.J.L., AlMazroa, M.A., Memish, Z.A., 2012. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380, 2163–2196. doi:10.1016/S0140-6736(12)61729-2

[199]. WebMD, 2014. Magnetic Resonance Imaging (MRI) [Online] WebMD. Available at: http://www.webmd.com/a-to-z-guides/magnetic-resonance-imaging-mri (accessed 8.1.15).

[200]. Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J., Hirschman, L., 2007. Rapidly retargetable approaches to de-identification in medical records. Proc AMIA Annu Fall Symp 14, 564–573. doi:10.1197/jamia.M2435

[201]. Wenham, C.Y.J., Grainger, A.J., Conaghan, P.G., 2014. The role of imaging modalities in the diagnosis, differential diagnosis and clinical assessment of peripheral joint osteoarthritis. Osteoarthr. Cartil. 22, 1692–1702. doi:10.1016/j.joca.2014.06.005

[202]. Westbrook, J.I., Coiera, E.W., Gosling, A.S., 2005. Do Online Information Retrieval Systems Help Experienced Clinicians Answer Clinical Questions? Proc AMIA Annu Fall Symp 12, 315–321. doi:10.1197/jamia.M1717

[203]. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A., 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 39, W541–5. doi:10.1093/nar/gkr469

[204]. Winkler, W.E., 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

[205]. Wintner, S., 2009. What Science Underlies Natural Language Engineering? Computational Linguistics 35, 641–644. doi:10.1162/coli.2009.35.4.35409

[206]. Woods, R.W., Eng, J., 2013. Evaluating the Completeness of RadLex in the Chest Radiography Domain. Academic Radiology 20, 1329–1333. doi:10.1016/j.acra.2013.08.011

[207]. Wu, C., Schwartz, J.-M., Nenadić, G., 2013. PathNER: a tool for systematic identification of biological pathway mentions in the literature. BMC Syst Biol 7 Suppl 3, S2. doi:10.1186/1752-0509-7-S3-S2

[208]. Wulff, H.R., 2004. The language of medicine. J R Soc Med 97, 187–188.

[209]. Yan, R., Wang, H., Yang, Z., Ji, Z.H., Guo, Y.M., 2011. Predicted probability of

meniscus tears: comparing history and physical examination with MRI. Swiss Med Wkly.

[210]. Yetisgen-Yildiz, M., Gunn, M.L., Xia, F., Payne, T.H., 2011. Automatic identification of critical follow-up recommendation sentences in radiology reports. Proc AMIA Symp 2011, 1593–1602.

[211]. Zeng, Q.T., Redd, D., Divita, G., Jarad, S., Brandt, C., Nebeker, J.R., 2011. Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. J Health Med Inform. doi:10.4172/2157-7420.S3-001

[212]. Zúñiga, G.L., 2001. Ontology: its transformation from philosophy to information systems, in: Presented at the Proceedings of the international conference on Formal Ontology in Information Systems, ACM, pp. 187–197. doi:10.1145/505168.505187