Chinh Nguyen Kim

A Text-based Ontology-driven Decision Support System

Master’s Thesis of Mathematical Information Technology

June 7, 2018

University of Jyväskylä

Faculty of Information Technologies

Author: Chinh Nguyen Kim

Contact information: chinhnk.93@gmail.com

Supervisors: Vagan Terziyan, Oleksiy Khriyenko and Pekka Neittaanmaki

Title: A Text-based Ontology-driven Decision Support System

Työn nimi: Tekstipohjainen ontologia-päätöksenteon tukijärjestelmä

Project: Master’s thesis

Study line: Web Intelligence and Service Engineering

Page count: 61 + 2

Abstract: The coming of the Big Data era has posed great challenges to the traditional decision support systems, which are unable to effectively leverage unstructured data, necessitating more flexible and adaptable approaches. Originating from the same acknowledgment expressed in the Value from Public Health Data with Cognitive Computing project, this study introduces a text-based approach to designing decision support systems and evaluates its practicality, utility as well as its advantages in facing these challenges. The potential benefits from leveraging Semantic Web technologies as a driving force and in improving the performance of such systems were also investigated. For assessing the validity of the approach in practice, two proof-of-concept prototypes were developed in succession.

Theoretical analysis showed that a text-based decision support system is fully capable of alleviating the difficulties faced by traditional systems in utilizing unstructured textual data in the decision-making process. On the other hand, the implementations of the prototypes demonstrated the possibility of employing large-scale and well-structured ontologies like SNOMED-CT as the basis for knowledge representation, resulting in performance gain. At the same time, the application of the proposed semantic relevance measure was shown to further enhance the derivation of relevant information. While additional and more conclusive evaluations are needed, the study proved that a text-based ontology-driven decision support system is feasible and worthy of further research.

Keywords: decision support systems, Semantic Web, ontology, text analytic, natural language processing, semantic similarity, semantic relevance

Suomenkielinen tiivistelmä: Abstract in Finnish…

Avainsanat: Keywords in Finnish…

Preface

It has been an unforgettable experience.

Writing these lines as I am finalizing the thesis, I am filled with memories coming back to me in waves. Looking back the last two-and-a-half years, starting from the moment I nervously sent the application documents to the University of Jyväskylä, the interview, the Letter of Acceptance, the joy… And then, came the first time I set foot on the plane to a foreign country, 7800 kilometers away from home. I am recalling the first courses of the first year, followed by an eventful summer with new friends and project works, which inspired this thesis study. The second year of study was harder as I have been working at the same time. I will never forget those struggling days and nights simultaneously trying to fulfill work requirements as well as writing this thesis.

During these times of hardship, I am ever thankful to Dr. Olena Kaikova, Prof. Vagan Terziyan and Dr. Oleksiy Khriyenko for their guidance, constant support and encouragement. Without them, it would have been impossible for me to complete this thesis on time, if at all. I would also like to dedicate special thanks to Prof. Pekka Neittaanmaki for facilitating the *Value from Public Health Data with Cognitive Computing* project, from which I gain much valuable knowledge and experience, as well as the ideas for my thesis.

My family and my significant other at home have always had faith in me and supported me. To them, a simple thank is never enough. I know I am blessed, and I am grateful.

In the end, I hope the research presented in this thesis will be of value, not only to me as a milestone of maturity, but also to others, as it may give ideas, inspire arguments, criticisms and further studies into the topic.

Best regards.

Jyväskylä, June 7, 2018

Chinh Nguyen Kim

Glossary

Decision support system

An information system that supports business or organizational decision-making activities.

Semantic Web

An extension of the World Wide Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". The Semantic Web is therefore regarded as an integrator across different content, information applications and systems.

Ontology

In computer science and information science, an ontology encompasses a representation, formal naming, and definition of the categories, properties, and relations of the concepts, data, and entities that substantiate one, many, or all domains.

Semantic Similarity

Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation (e.g. their string format).

List of Figures

Contents

1 Introduction

As early as the dawn of software systems in the 1950s and 1960s, it was envisioned that computers and their impressive processing power over a large quantity of data could be involved in assisting humans in the process of decision making (Keen, 1978), thus began the research on decision support systems. Over the years, the focuses and definitions of decision support system have evolved with the introduction of new technologies and data sources. However, two common characteristics of most decision support systems have remained largely unchanged: the high complexity and rigidness of the center data model and the limited capability in processing unstructured data. In the past, these shortcomings were not always apparent since, in practice, most decision support systems were implemented to deal with problems where data scope and structure were well-defined or of specific domains with precise objectives. In today's context, however, with the speed, volume, complexity, variety and variability of data growing ever so quickly (Gantz & Reinsel, 2012), these drawbacks are amplified and can no longer be ignored. Power et al. (Power, 2014) outlined several key challenges for decision support systems in this so-called *Big Data* era, among which is the processing of semi-structured and unstructured data, which are mostly made up of textual artefacts (Holzinger et al., 2015).

The aforementioned issues were also of the main concerns expressed by doctors – the primary users of various clinical decision support systems – and other stakeholders during the *Value from Public Health Data with Cognitive Computing*[1] project. The project was a cooperation between IBM, the University of Jyväskylä, Central-Finland Central Hospital (KSSHP) and experts in both technology and medical domain, with the mission of exploring potential applications of cognitive technologies in order to improve healthcare quality in the highly-digitalized Finland. Throughout the workshops and follow-up discussions, one of the emerging patterns was that currently employed clinical software systems, mainly electronic medical record (EMR) management systems, while often large and complex, lack high-level supporting functionalities. The inability to interpret and meaningfully aggerate large-scale

[1] https://www.jyu.fi/it/fi/tutkimus/tutkimushankkeet/paattyneet-hankkeet/tekes/publichealthdata

medical data, with text-based patient health records being the most prevalent, to assist doctors in making diagnoses and treatment plans was regarded as the main limiting factor affecting the utility of such systems in practice.

Based on these observations from both literatures and feedbacks from domain experts, we believe that the addition of a more flexible, text-focused approach to designing decision support systems is an objective necessity.

Simultaneously with the trend of Big Data, the *Semantic Web* (Berners-Lee, Hendler, & Lassila, 2001) and its collection of ontologies and supporting technologies have significantly expanded both in size as well as sophistication. In addition to enabling the contextualization of millions of web-based contents, the network of linked data which has its foundation grounded on Semantic technologies has contributed greatly to many domains of research and industry, most notably biology and medicine. In addition, large-scale ontologies such as *WordNet* (Miller, 1995), *DbPedia* (Auer et al., 2007) and *YAGO* (Suchanek, Kasneci, & Weikum, 2008) have seen recent uses as reference sources for various text processing applications. While usually utilized as vocabulary sets, the rich and extensive structural information of these ontologies has not been effectively exploited by such applications.

In this work, we explore an alternative to traditional designs of decision support systems, one that not only is able to overcome the challenges posed by the unstructured big data, but also employ these sources as the basis for decision-making, through both theoretical analysis and practical implementation of proof-of-concepts. The main research question of the thesis thus is as follow:

- ***How to support a human decision-making process on the basis of unstructured big data?***

We also attempt to address two additional support questions in the process of finding and evaluating the answer to the first, which are:

- ***How to formalize and implement the concept of "relevance" in data filtering within the context of a text-based decision support?***

- ***How to boost the performance of a text-based decision-support system using semantic technologies?***

1.1 Research Methodology

Since this thesis study originated from within the scope of a project aiming at generating solutions for real-life problems, we followed the ***Design Science Research Method (DSRM)*** (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), which consists of six activities graphically described in Figure 1.

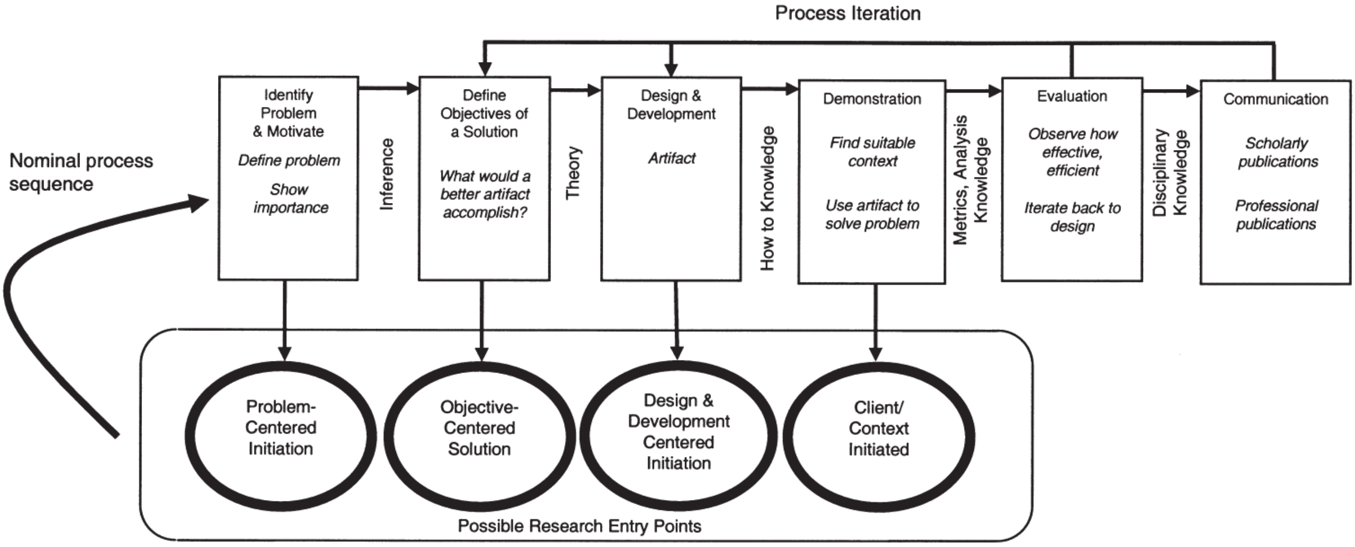

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 1. DSRM Process Model

(Peffers et al., 2007)

The first two activities in the research process, namely *Problem identification and motivation* and *Define the objectives for a solution* were carried out in a series of workshops at the beginning of the Value from Public Health Data with Cognitive Computing project. During these workshops, we collaborated with technical and medical experts in order to discover obstacles in healthcare practices which can be overcome by the application of cognitive technologies. The result of the workshops consists of a collection of use-cases and user-stories, from which a handful were selected for further assessments and solution prototyping based on feasibility and criticality. Among the selected use-cases, we focused on the *Driving Assessment* use-case, in which a text-based decision support system can be constructed to assist doctors in evaluating a patient's medical records with respect to a set of regulations in order to determine whether the patient meets the health requirements for a driving license.

The *Design and development* activity in our research was characterized by a systematic review of applicable cognitive computing and artificial intelligence technologies, including IBM products, followed by the iterative development of two prototypes. The first prototype was demonstrated on various occasions to stakeholders from whom we received feedback (*Demonstration* and *Evaluation*). While the majority of the feedback was positive and the work was published in (Khriyenko, Nguyen Kim, & Ahapainen, 2018) (*Communication*), we were not satisfied with the performance of the system and developed a second prototype. This prototype was demonstrated to outperform its predecessor with added functionalities and increased utility.

1.2 Thesis Structure

We divided the composition of the thesis into six main parts, excluding this introduction. Chapter 2 presents the case for the text-based approach to designing decision support systems by first addressing the difficulties faced by traditional decision support systems followed by general descriptions and arguments in favor of the new approach through retrospective analysis. The applicability of the ontology-driven method for a text-based decision support system is discussed in Chapter 3. In this chapter, we also propose a new way of incorporating semantic similarity measures in text analysis, attempt to resolve the *similarity-relevance* debate in the context of ontology-based analyses and outline the general architecture of a text-based ontology-driven decision support system. In Chapter 4, we document the development of our two prototypes of the proposed decision support system with the specific use-case of assisting a doctor in in assessing a patient's fitness to drive. The evaluation of these prototypes is provided in Chapter 5, including their limitations and our plans for future improvements. A number of related works are discussed in Chapter 6 and the conclusions of the thesis are given in Chapter 7.

2 Text-based Decision Support Systems

2.1 Overview of Traditional Decision Support Systems

A ***decision support system*** (DSS) has been broadly defined as any system that facilitates and increases the efficiency of the decision-making process, including software systems (Keen, 1980). The purposes, structures and implementations of such systems vary, but according to (Power, 2002), they can generally be categorized as: *communication-driven DSS, data-driven DSS*, *model-driven DSS, document-driven DSS*, and *knowledge-driven DSS*.

- A ***communication-driven DSS*** focuses on enabling networking and cooperation between a group of decision makers. This board definition includes any modern communication software applications such as Skype[2] or Google Hangouts[3], or any collaboration platforms (e.g Atlassian Confluence[4]), and even tools with integrated collaboration features such as Google Docs[5].
- A ***data-driven DSS*** centers around the acquirement and manipulation of data, both internal and external, in the format of time series. This kind of DSS aims at providing decision makers with clearer insights which serves as the basis for decisions.
- A ***model-driven DSS*** assists decision makers in assessing a situation by running provided data through a central model configurable via adjustable parameters. The model itself is constructed to reflect certain statistical aspects, to solve optimization problems or serve as a simulation. An open source example of a DSS of this kind is Dicodess (Gachet, 2004).
- A ***document-driven DSS*** deals with unstructured information stored in electronic formats. Its functionalities mainly concern managing these assets and retrieving them upon being queried.

[2] https://www.skype.com/
[3] https://hangouts.google.com/
[4] https://www.atlassian.com/software/confluence
[5] https://www.google.com/docs/about/

- A ***knowledge-driven DSS*** relies on a predefined knowledge-base to drive the decision-making process. Within this knowledge-base there is stored problem-solving expertise structured as a set of facts, rules or some sort of procedures.

As important as the role of a cooperation platform is, we decided not to consider communication-driven DSS partly because this topic has always been spearheaded by enterprise solutions almost to the point of saturation and mostly because it is not among the main focuses of our studies.

There are certain inherent disadvantages of data-driven DSS-es and model-driven DSS-es, however, the first of which is the fact that they can only make use of highly structured data sources. In addition to limiting the range and volume of data which could be used in decision making, this also requires considerable additional efforts to be invested into data preparation and pre-processing. The second disadvantage of data-driven DSS-es is that, as more data dimensions are added, the analytic models tend to rapidly expand in terms of complexity. Moreover, and consequentially, the rigid nature of the DSS makes horizontal scaling to address similar problems extremely difficult if not outright impossible which necessitates redesigning the whole system.

The traditional approaches with document-driven DSS-es where documents are mostly stored and queried using information retrieval techniques stem from the great challenges to truly *"understand"* and meaningfully process unstructured textual data. Though still bringing about added value, these approaches leave much to be desired.

The knowledge-driven approach for DSS-es generally provides more versatility to be applied in solving real-life problems, especially in conjunction with the document-driven method. However, knowledge-bases require significant manual efforts to construct and maintain, with extensive involvement of field experts in curation – a resource that can be inaccessible at times. Moreover, due to the work invested in building a knowledge base, it is often customized to a specific set of problems for maximum effectiveness, thus sharing the same disadvantage of model-driven DSS-es in terms of scalability.

2.2 The Case for Text-based Decision Support Systems

The recent years have seen an exponential expansion of data, partly due to the massive increase in the complexity of our society and the reality that nowadays almost every person is a potential data creator. While the volume of the whole digital universe was projected to grow by a factor of 300 from 2005 to 2020 and reach the staggering number of 40 trillion gigabytes by International Data Corporation (IDC) (Gantz & Reinsel, 2012), 70-80 percent of its data consists of unstructured contents, as estimated in 2013 by the Computer World magazine (Holzinger et al., 2015). In turn, making up the vast majority of these unstructured data are text-heavy contents, ranging from digitalized books, journals, legal documents, health records to email, messages and Web pages. All this abundance and richness poses both great challenges and near limitless opportunities.

In facing the challenges to DSS-es in the Big Data era, we believe it is important lay emphasis on the ability of a DSS to make the most of existing materials and its flexibility in integrating future contents. We thus propose a fully text-based design in building a DSS where its data foundation consists fundamentally of human-readable documents and its *"knowledges"* are inferred directly from said texts. When compared with traditional approaches, a text-based DSS holds absolute advantage over its data- and model-driven counterparts due to the latter's inability to effectively process unstructured data. Such a DSS is also capable of much greater utility than a document-driven DSS thanks to its ability to reason and extract knowledge from provided textual data, which also gives the text-based DSS an edge over a knowledge-driven DSS as it eliminates the need for a laborious knowledge curation process. The capability to derive knowledge from text is discussed more in-depth in Chapter 3.

Additionally, a text-based approach may well make it possible to access new niches where the application of traditional DSS-es has been minimal. A notable example of these is the legal system and its many fields of law, which have become more and more ubiquitous in the modern society. Other examples include industries in which the integrity of practice can only be guaranteed through heavy regulations such as banking and finance or and those where business models rely directly on applying complex sets of policies such as insurance. The common characteristic of these niches is the need to correctly and efficiently process large amounts of human-targeted regulative documents and also apply them to documented cases. Not only do the volume and complexity of these rules make them almost impossible to be formally modeled, their dynamic nature can only be coped with by the flexibility offered by a text-based solution.

2.2.1 Document Types

Our proposition of a text-based DSS arose from the observation that a body of text being of concern in the decision-making process is generally either descriptive (declarative) or regulatory (imperative), or a mixture of both. In the latter case, such a text can be subsequently broken down into smaller bodies of text which belong to either category. Hence, we define ***base documents*** as collections of text where imperative contents are prevalent and ***target documents*** as those whose main focus is the descriptions of objects and phenomena. As the names suggest, in a text-based DSS, base documents provide the grounds for the decision-making process while contents of the target documents will be assessed in regards of said grounds.

2.2.2 General Use Cases

In practice, the design of our text-based DSS suits two general use-cases, which can be adapted to be used in many of fields and industries where textual materials are common in the form of regulations, contracts, reports or claims.

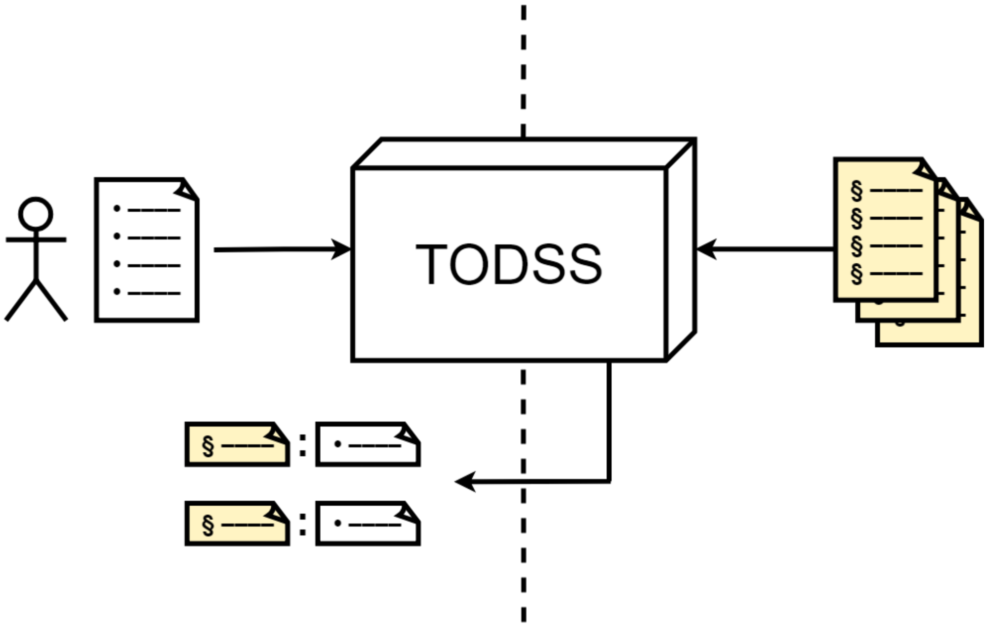


Figure 2. First use-case: Descriptions → Concerning rules

The first use-case involves having a set of target documents with descriptions of an entity (e.g. human, object, event, etc.) as input. A set of base documents has already existed within the DSS, providing reference rules. Upon consultation, the DSS will return suggested courses of action with rationales presented in the narrowed collection of rules applying to given descriptions (see Figure 2).

In real life, this use-case represents any situation where the decisions must conform to a set of complex, intertwined and, sometimes, conflicting rules. A primary example of this is navigating the legal landscape when an entire law can be stored within the DSS without the need of heavy preprocessing and used as reference in assessing materials such as text-based evidence and transcribed testimonies. Similarly, a potential premium holder can easily seek assistance of the DSS to see what terms and conditions might be applied to his or her described situation.

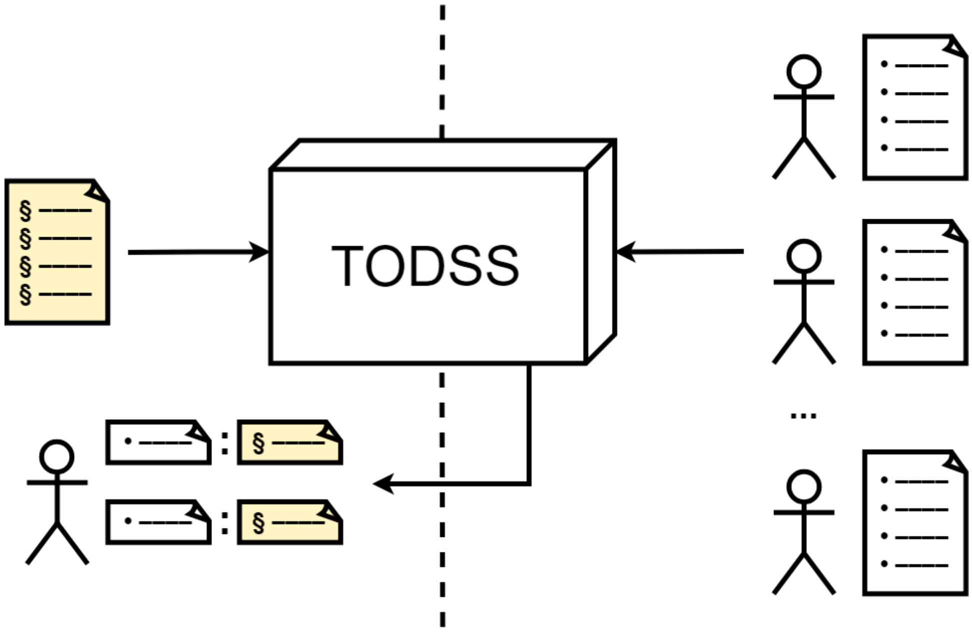


Figure 3. Second use-case: Rules → Matching descriptions

To a certain degree, the other use-case can be considered a reversal of the first one, as shown in Figure 3. In this use-case, the decision maker is in possession of a base document and wishes to find out which candidates are to be affected by the rules specified in said document. The DSS will then take the base document as input and output a filtered list of candidates whose descriptions are covered by the new rules.

In practice, the second use-case can materialize into any situation in which there exists a database of entities that are subjected to some sort of regulations or conditions and have textual descriptions. These entities can be persons, companies, contracts or business plans, etc. By further analyzing the focused list, the decision maker can decide whether to go forward with the proposed legislation or to take actions on the listed individuals according to the newly given rules.

Being domain independent and not requiring heavy human involvement, the two described use-cases exhibit a high degree of generalizability. Most real-life applications will likely see both use-cases mixed or scaled to some extent due to the need for exploring data from different angles and in varied scopes. Nevertheless, it can be easily achieved thanks to greatly simplified data requirements.

3 The Ontology-driven Approach

In our proposal, the ***ontology-driven*** approach to designing a DSS refers to the extensive use of information from ***Semantic-Web***-compliant ***ontologies*** to discover and represent knowledges from texts, which, in turn, forms the basis for decision making.

3.1 The Semantic Web and Ontologies

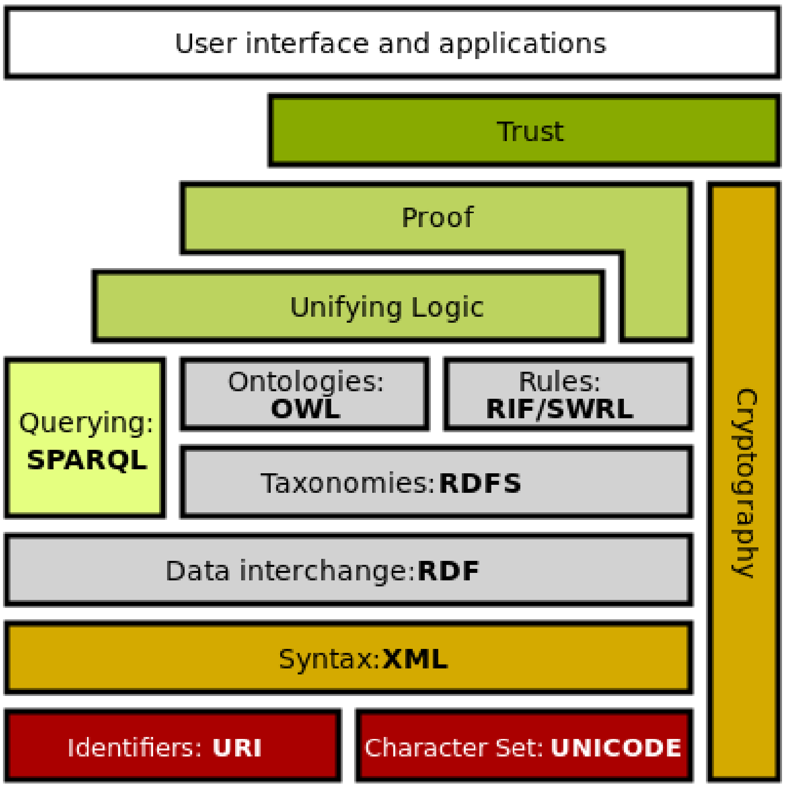


Figure 4. The Semantic Web Stack

Source: https://en.wikipedia.org/wiki/Semantic_Web_Stack

Envisioned to be an extension to the World Wide Web, the ***Semantic Web*** (Berners-Lee et al., 2001) has been adopted and extensively developed by the World Wide Web Consortium (W3C). The goal of this *"Web of Data"* is to serve as a medium for the automatic creation, distribution and context-aware utilization of linked data, through a series of standards and technologies such as RDF (Klyne & Carroll, 2006), SPARQL (Prud & Seaborne, 2006), RDFS[6] and OWL (Bechhofer, 2009). Over the years, additional standards and technologies have been introduced to improve the utility and versatility of the Semantic Web, such as SKOS (Isaac & Summers, 2009), JSON-LD (Sporny, Longley, Kellogg, Lanthaler, & Lindström, 2014) and SWRL (Horrocks et al., 2004). Altogether, they make up the so-called *Semantic Web stack* (Figure 4).

An ***ontology*** in the Semantic Web is roughly synonymous with a *vocabulary* used in an application. In this vocabulary, the terms are effectively defined along with their characteristics, taxonomy and possible relationships. An ontology can be very simple (consisting of only one or two concepts) or very complex (describing several thousand interconnected terms).

In practice, many ontologies have been constructed to serve a variety of purposes, from specifying friendships and relationships (FOAF[7], Fedora (Lagoze, Payette, Shin, & Wilper, 2006)), facilitating data sharing and integration (SKOS, UMBEL[8]), organizing glossaries for a specific field of scientific studies (NASA Thesaurus (NASA STI Program, 2012), GeoNames[9], the Music Ontology (Raimond, Abdallah, Sandler, & Giasson, 2007), UMLS (Bodenreider, 2004)), to storing extracted general knowledge (Dbpedia (Auer et al., 2007), YAGO (Suchanek et al., 2008)) or serving as basis for various natural language processing applications (WordNet (Miller, 1995)). Many of these ontologies have been constantly growing and contain general and domain-specific data in hundreds of thousands of classes and millions of data entries. It is without questions that they are among the most extensive and valuable data sources made available for any kind of application.

3.2 Ontology-based Entity Annotation

In the field of natural language processing, *named entity recognition* has long been studied and used to extract information from text bodies. *Named entities* refer to those real-world

[6] https://www.w3.org/TR/rdf-schema/
[7] http://xmlns.com/foaf/spec/
[8] http://umbel.org/resources/about/
[9] http://www.geonames.org/ontology/documentation.html#

objects such as humans, organizations, locations, products etc. which hold enough significance to be individually named. Abstract objects can also be considered named entities (e.g. *"the Internet"*).

Annotation of named entities is the process of marking out tokens like words or phrases representing the entities as they appear in texts. For example, the sentence:

The current President of the United States, Donald J. Trump, withdrew
the country from the Paris Agreement.

after annotation, might looks like:

The current [President of the [United States] $_{\text{Nation}}$] $_{\text{Position}}$, [Donald J.
Trump] $_{\text{Person}}$, *withdrew the country from the* [Paris Agreement] $_{\text{Treaty}}$.

with each block being an entity. The annotated texts are normally stored for future uses using markup languages like XML or in data structures defined by frameworks such as Apache UIMA (Ferrucci & Lally, 2004).

While identifying named entities is a trivial task for a relatively knowledgeable person, automating it requires either detailed grammar-based solutions or sophisticated large-scale machine-learning models due to both the lack of a universal library of everything and the overwhelming amount of ambiguity in the *"general"* context. Within the Semantic Web, however, these fundamental problems are virtually alleviated. On one hand, every *"thing"* in the Semantic Web can be uniquely identified and referenced via an URI, thus making the web and its massive collection of ontologies a *"library of everything"*. On the other hand, an ontology can dictate the context in which a text is understood by, for example, restricting the terms that of concern during the analysis and what meaning they carry (i.e. disambiguation).

Using ontologies as a basis for entity annotation also comes with added benefits. To begin with, annotated data can be linked directly to external resources across the Semantic Web and contribute to a vast network of linked data. This opens up the possibility to cross-referencing and semantic inferences, which further improve data value. Moreover, since ontologies, by their nature, transcend the language barrier, solutions in which analyses relying predominantly on ontology-based entities can be scaled to other languages with little effort.

3.3 Semantic Similarity

3.3.1 Concept-level

Semantic similarity measures the likeness between two concepts with respect to their meaning instead of their lexical resemblance to each other. For example, the term *"bank"* would have much higher similarity score to *"credit union"* than *"tank"*. The metric brings considerable added value to analyses based on natural language processing as it enables the assessment of terms beyond their exact-match surface forms.

When it comes to formularizing estimations for semantic similarity, there have been two main approaches. The first aims at using the characteristics of the topology (i.e. the ontology) and the positioning of the terms in said structure, while the second learns the scores from statistical models based on large text corpuses where presences of the terms in questions can be found. In this thesis, we focus on the approach that exclusively utilizes ontologies due to the uncertainty in having access to an ideal domain corpus in practice – one with exhaustive coverage of domain concepts while still maintaining a balanced distribution of said terms.

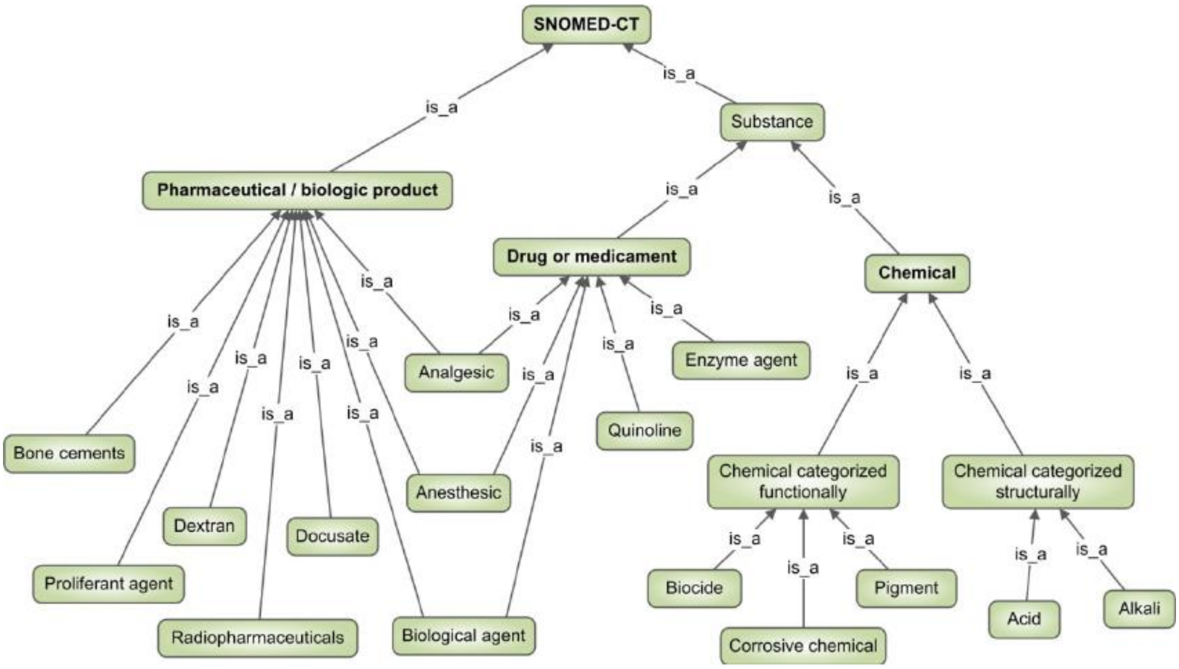


Figure 5. Part of the class hierarchy of SNOMED-CT ontology

Source: (Gómez-Pérez, Martínez-Romero, Rodríguez-González, Vázquez, & Vázquez-Naya, 2013)

As ontologies are technically graphs of data where nodes are made up of classes and individuals and edges denote relationships (Figure 5), many ontology-based semantic similarity measures can be effectively categorized as either *edge-based* or *node-based*. Most of node-based measures, however, starting with (Resnik, 1995) and later developed upon in (Jiang & Conrath, 1997), (Lin, 1998) and (Maguitman, Menczer, Roinestad, & Vespignani, 2005), require the *information content* (IC) of a term to be determined. IC of a concept is formally defined as the inversed log likelihood of that concept appearing in a corpus, resulting in less frequent terms being considered more informative. Not only this does approach entail expensive pre-calculations, it also shares the same potential impracticality with the statistical methods.

Among the *edge-based* measures, the most straightforward method uses the shortest path length between the concepts as their semantic distance (Rada, Mili, Bicknell, & Blettner, 1989):

$$dis_{PL}(c_1, c_2) = \min(number\ of\ edges\ connecting\ c_1\ and c_2)$$

Since only *is-a/subClassOf* edges hold categorical information, this formula can be rephrased as the sum of path lengths from each concept to their *least common subsumer* (LCS). Various improvements to this measure, in which additional features of the taxonomy were considered, were proposed. These features include, for example, the depth of the LCS (Wu & Palmer, 1994), the maximum depth of the ontology (Leacock & Chodorow, 1998), or the level difference between the two concepts (Choi & Kim, 2003). Others even went with more sophisticated approaches, from applying optimization to coefficients of an exotic non-linear function (Al-Mubaid & Nguyen, 2006) to a cluster-based measure (Al-Mubaid & Nguyen, 2006).

While the aforementioned measures eliminate the need for large text corpuses and are computationally more efficient than the node-based and statistical counterparts since the scores can be calculated on-the-fly, they are not without drawbacks. In addition to being heavily dependent on the quality of the ontology in order to get a good measurement, most of them assess semantic similarity basing solely on *minimum path* between concepts. In ontologies where multiple-inheritance is commonplace (e.g. SNOMED-CT, Dbpedia), this problem becomes apparent since a significant portion of categorical information is simply ignored.

Similar conclusions were also shared by Batet et al (Batet, Sánchez, & Valls, 2011), who, in turn, proposed their own semantic similarity measure:

$$sim_{BSV}(c_1, c_2) = \log \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|}$$

where $T(c_i) \ = \{c_j \in C \mid c_j \ is \ superconcept \ of \ c_i\} \ \cup \ \{c_i\}$ with C being the set of all concepts within the ontology. This set-based measure does indeed circumvent the shortcomings of existing methods and is on-par with edge-based measures in term of computational cost. It is worth noting, however, the codomain of the function ranges from *0* to ∞. Even though most of the values in practice rarely exceed *20* and small values (between *1* and *4*) have been shown in experiments to largely correspond with experts' ratings, it is not suitable to be used in text analysis applications without normalization.

From that observation and meanwhile still maintaining acknowledgement of the advantages of the set-based approach, we finally decided to fall back on a classical similarity measure – the *Jaccard Index*:

$$sim(c_1, c_2) = \frac{|T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} = \frac{|T(c_1) \cap T(c_2)|}{|T(c_1)| + |T(c_2)| + |T(c_1) \cap T(c_2)|}$$

where $T(c_i)$ is defined analogously with that in the proposal of Batet et al. With the semantic similarity scores of a pair of concepts then fall within *0* and *1*, we proceeded to formulate a way to compute semantic similarity on higher levels.

3.3.2 Document-level

For estimating the relatedness between two text bodies of *arbitrary length* (from now on referred to as *documents*), we propose a two-step approach in which the basic units for calculations are concept-level similarity scores. The main advantage of this approach is the ability to introduce contexts into text similarity analysis, which has been predominantly relying on plain statistical models (e.g. Apache Solr scoring functions[10]).

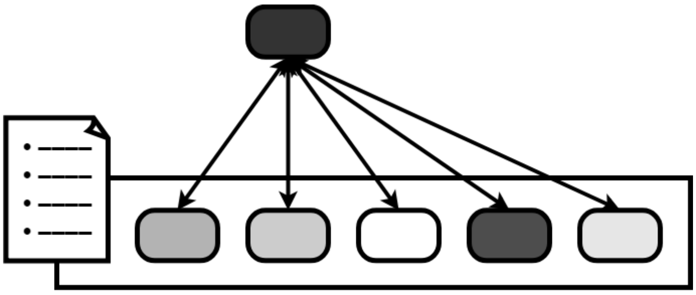


Figure 6. Concept-Document similarity

The first step is determining the relevance of a term or concept to a document. With the assumption that the collection of all ontology-based entities found in the document adequately represents the document itself within the considered context, the similarity between a concept and a document can be effectively estimated as some aggregation of the semantic similarity between said concept and individual document entities (Figure 6). However, it is

[10] http://lucene.apache.org/core/3_5_0/api/core/org/apache/lucene/search/Similarity.html

also important to take into account the significance of each distinct entity to its encapsulating text. Thus, we define the *semantic relevance between a concept c and a document d* - $sim(c,d)$ as:

$$rel(c,d) = \sum_{i=1}^{n} sim(c, c_{di}) \times p(c_{di})$$

where c_{di} is an ontology-based entity mentioned in d and $p(c_{di})$ is the probability of c_{di} appearing in d, which itself is calculated as:

$$p(c_{di}) = \frac{count(c_{di})}{|\{c \in C | c \; is \; mentioned \; in \; d\}|}$$

It is worth noting that the same measurement can be applied when concept c is also mentioned in d.

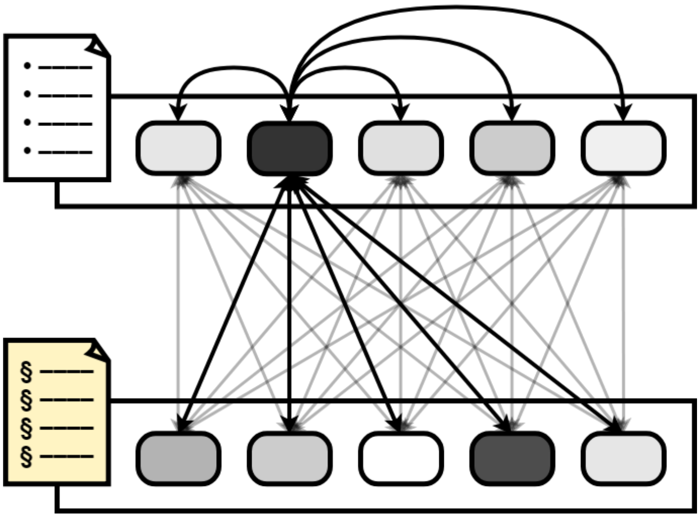


Figure 7. Document-document relevance

Likewise, the second step is a scale-up from the first one. In this step, the semantic similarity between two documents is assessed using the newly defined concept-document similarity measure. We consider the relevance connections between an entity and each document to indicate the similarity between the two documents with respect to the topical aspect represented by that entity (Figure 7). Thus, the *semantic similarity between two documents* d_1 *and* d_2 can be formally given as:

$$sim(d_1, d_2) = \sum_{j=1}^{m} rel(c_j, d_1) \times rel(c_j, d_2) \times p(c_{di})$$

where c_{dj} is a concept in d_1. The measurement function is bidirectional and has a codomain which also falls between *0* and *1*.

3.4 Relevance in a Text-based DSS

In information retrieval, the terms “relevance” and “similarity” have always been closely associated to the point that similarity measures between the features of the query and those of the result are normally used to reflect the degree of relevance. The *cluster hypothesis* (Jardine & van Rijsbergen, 1971) suggested that documents belonging to the same cluster (i.e. similar to each other) have a high likelihood to share the same level of relevance with respect to the information needs. Search engines, such as Google, also make the same assumption in ranking search results, as demonstrated in the *Google similarity distance* (Cilibrasi & Vitanyi, 2007).

However, it has also been argued that similarity can only represent *topical relevance* at best, not *user relevance*. Already at beginning days of information retrieval, W. Cooper pointed out that there is a distinction between *logical relevance* and the *utility* of an information retrieval system (Cooper, 1971). Over the years, attempts have been made to address this gap, including evaluating documents with several criteria such as aboutness, coverage, appropriateness, and reliability (da Costa Pereira, Dragoni, & Pasi, 2012) and even redefining *“relevance”* to be *“situational”* (Borlund, 2003).

In our text-based DSS, while the purpose of a *target document* is to describe and the purpose of a *base document* is to regulate, it is perfectly possible for them to share a common topic or, in other words, be *topically similar*. As discussed in 3.3.2, this similarity is reflected by *document-level semantic similarity* measure. Additionally, in the proposed general use-cases, a query is usually the whole document of one type and the information needs in such cases are for extracts of those of the other type. Thus, the semantic similarity between two documents can be considered to indicate both *topical relevance* as well as *user relevance* within the context of the DSS.

3.5 General Architecture

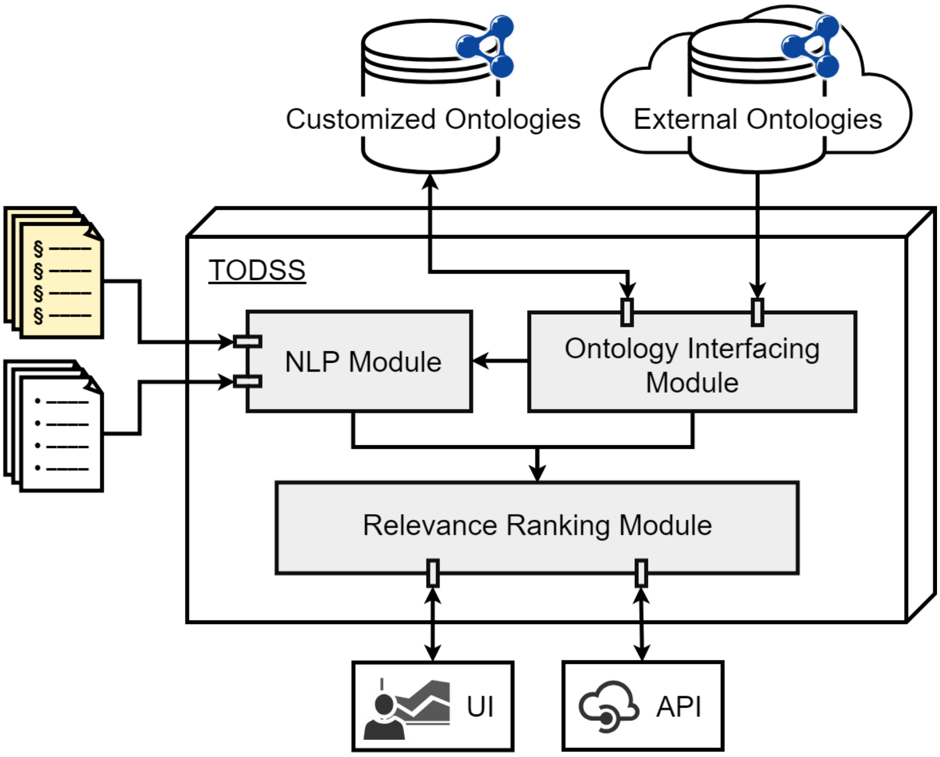


Figure 8. TODSS general architecture

From the architectural point of view, a Text-based Ontology-driven Decision Support System (TODSS) consists of three main internal components: the *NLP module*, the *Ontology Interfacing* module and the *Relevance Ranking* module. The *Ontology Interfacing* module is responsible for retrieving, parsing and storing ontological data from external sources, as well as providing the capability to import custom-defined ontologies. The module also provides the *NLP* module with vocabulary sets from which to annotate the documents post segmentation. The annotated document sections together with taxonomical data from stored ontologies then serve as input for the *Relevance Ranking* module, which conducts analyses upon the user's requests. Alternatively, interactions with the system can also be performed via an

Application Programming Interface (API). An illustration of the general architecture of a TODSS is given in Figure 8.

At an abstract level, in our text-based DSS, documents of both types make up the system *database* with factual knowledge largely contained within *target documents* and rules specified in *base documents*, while the decision *contexts* are determined by the ontologies in use.

4 Development of Proof-of-Concept Prototypes

In order to assess the feasibility of the theoretical design of a TODSS, we implemented a series of proof-of-concept prototypes. The first prototype was developed within the scope of the Value from Public Health Data with Cognitive Computing project. With the specific use-case of assisting doctors in evaluating patients' fitness to drive, we aimed at developing a DSS that analyzes textual content of medical records for risk indicators with respect to a guideline document which was also given in natural language. The use-case was formulated from the real-life needs of doctors who have been increasingly overwhelmed by the amount of data requiring consideration within a limited timeframe. The second prototype was developed shortly afterwards as an incremental improvement. In this prototype, we focused on enabling the interaction with the system via APIs, formalizing its ontology storage for better integration with the Semantic Web and boosting the robustness through the implementation of the document ranking mechanism based on semantic similarities.

4.1 Data Sources

Both of our prototypes use the same data sources due to limited access to real-life medical data and the assumption that simulated data created by non-experts would bring about biases and inaccuracies.

4.1.1 Base Document

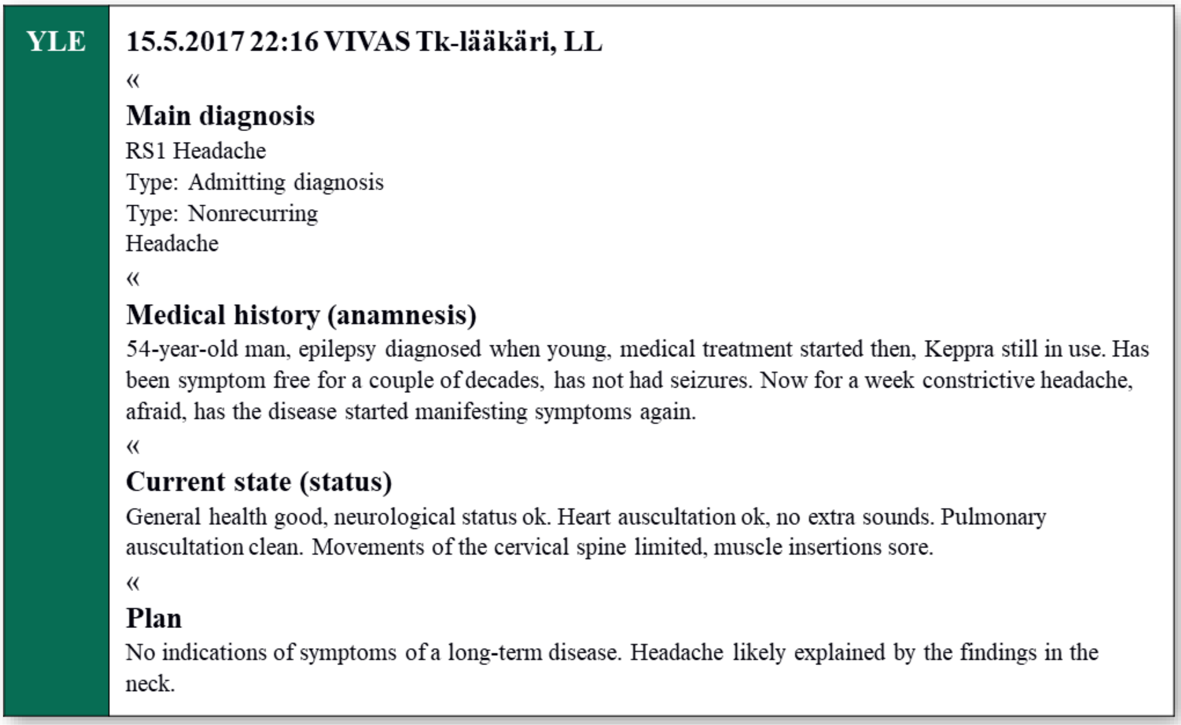
Since the first prototype was of a system which aids doctors in assessing whether a patient meets the health requirements to drive a vehicle (i.e. does not have any medical condition which might interfere with driving), we chose the Finnish *Ajoterveyden arviointiohjeet lääkäreille*[11] (transl. *Assessing fitness to drive: guide for medical professionals*) as our base document. However, since the support for processing the Finnish language was inadequate at the time, we opted for a translated English version of the document. Due to time and

[11] https://www.trafi.fi/liikennejarjestelma/liikenne_ja_terveys/tieliikenne_ja_terveys/ajoterveysohjeet_laakarille

resource constraints, only sections of the documents concerning medical conditions were translated.

4.1.2 Target Documents

We were given 21 anonymized doctor notes as our target documents for analysis. These notes are real-life data extracted from an *Electronic Medical Record System* currently in use, each records the outcome of a patient visit to the hospital. Any information that may lead to the identification of the patient was omitted from the notes. Only textual contents were considered. An example of the doctor notes is given in Figure 9.



YLE

15.5.2017 22:16 VIVAS Tk-lääkäri, LL

«

Main diagnosis

RS1 Headache
Type: Admitting diagnosis
Type: Nonrecurring
Headache

«

Medical history (anamnesis)

54-year-old man, epilepsy diagnosed when young, medical treatment started then, Keppra still in use. Has been symptom free for a couple of decades, has not had seizures. Now for a week constrictive headache, afraid, has the disease started manifesting symptoms again.

«

Current state (status)

General health good, neurological status ok. Heart auscultation ok, no extra sounds. Pulmonary auscultation clean. Movements of the cervical spine limited, muscle insertions sore.

«

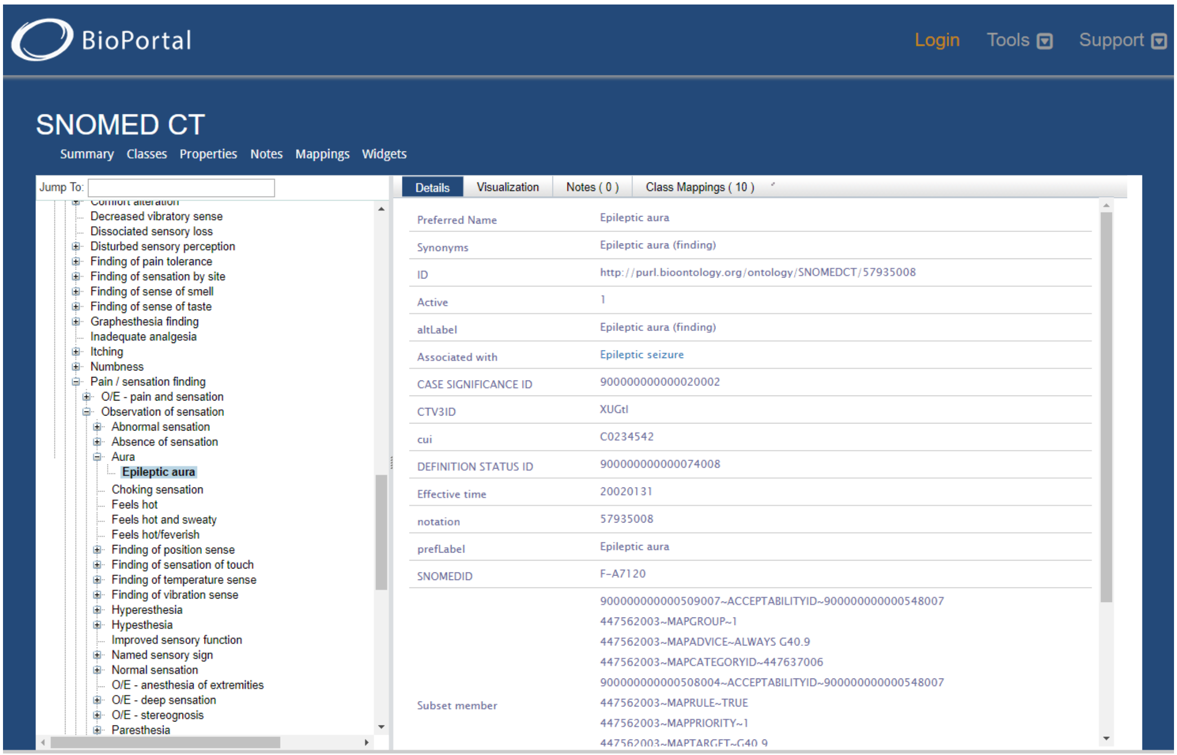
Plan

No indications of symptoms of a long-term disease. Headache likely explained by the findings in the neck.

Figure 9. Example of translated doctor notes

4.1.3 Ontology

As the domain of both prototypes is medical and health care, we chose *SNOMED CT* as our reference ontology for the analysis. *SNOMED Clinical Terms* (Donnelly, 2006) is an extensive and comprehensive clinical terminology which provides codes, synonyms and description of medical concepts. It is adopted as a standard for electronic health information exchange by the United States and the United Kingdom and is used by more than fifty countries worldwide. Compared to other alternatives such as *Medical Subject Headings* (Lipscomb, 2000) or ontology representations of *ICD (International Classification of Diseases)*[12], SNOMED CT offers a much larger number of classes with superior structural information.



- Comfort alteration
- Decreased vibratory sense
- Dissociated sensory loss
- Disturbed sensory perception
- Finding of pain tolerance
- Finding of sensation by site
- Finding of sense of smell
- Finding of sense of taste
- Graphesthesia finding
- Inadequate analgesia
- Itching
- Numbness
- Pain / sensation finding
  - O/E - pain and sensation
  - Observation of sensation
    - Abnormal sensation
    - Absence of sensation
    - Aura
      - **Epileptic aura**
    - Choking sensation
    - Feels hot
    - Feels hot and sweaty
    - Feels hot/feverish
    - Finding of position sense
    - Finding of sensation of touch
    - Finding of temperature sense
    - Finding of vibration sense
    - Hyperesthesia
    - Hypesthesia
    - Improved sensory function
    - Named sensory sign
    - Normal sensation
    - O/E - anesthesia of extremities
    - O/E - deep sensation
    - O/E - stereognosis
    - Paresthesia

Preferred Name	Epileptic aura
Synonyms	Epileptic aura (finding)
ID	http://purl.bioontology.org/ontology/SNOMEDCT/57935008
Active	1
altLabel	Epileptic aura (finding)
Associated with	Epileptic seizure
CASE SIGNIFICANCE ID	900000000000020002
CTV3ID	XUGtl
cui	C0234542
DEFINITION STATUS ID	900000000000074008
Effective time	20020131
notation	57935008
prefLabel	Epileptic aura
SNOMEDID	F-A7120
Subset member	900000000000509007~ACCEPTABILITYID~900000000000548007 447562003~MAPGROUP~1 447562003~MAPADVICE~ALWAYS G40.9 447562003~MAPCATEGORYID~447637006 900000000000508004~ACCEPTABILITYID~900000000000548007 447562003~MAPRULE~TRUE 447562003~MAPPRIORITY~1 447562003~MAPTARGET~G40.9

Figure 10. Example of a SNOMED-CT ontology class on BioPortal

BioPortal (Noy et al., 2009) is an integration of ontology services and related tools by The National Center for Biomedical Ontology, funded by the US National Institutes of Health. The service houses a collection of 593 aligned ontologies and supports various access methods, including a web interface for browsing through ontologies (Figure 10) and a SPARQL

[12] http://www.who.int/classifications/icd

endpoint and a set of REST APIs which can also be used for queries. Among these, the REST APIs proved to be the most robust and versatile, thus selected for use in our solutions. The SNOMED-CT ontology available on BioPoral is an RDF representation of SNOMED-CT, containing *327128* classes and *152* property types available here. Since the main focus of our solution is to detect and conduct analysis based on medical concepts, we first made use of *108056* subclasses of *ClinicalFinding*.

4.2 The Uses of IBM Watson Services

In our prototypes, we largely relied on two IBM Watson services for NLP-related tasks. This decision was grounded on two reasons: first, the IBM services were made conveniently available for us for research and evaluation purposes and second, we did not wish to dedicate too much time and effort on NLP tasks but instead on implementing analyses based on semantic similarity measures.

4.2.1 IBM Watson Natural Language Understanding

The first service we used was *IBM Watson Natural Language Understanding (Watson NLU)*[13] one of the cognitive computing services available on *IBM Cloud*[14]. The service provides the capability to analyze a text for a variety of language-based features including *categories, concepts, emotion, entities, keywords, metadata, relations, semantic roles* and *sentiment*. These analyses can be performed using either the built-in language model for general domain which can categorize documents into 1083 categories and recognize up to 24 entity types, 433 entity subtypes and 53 relation types or a more domain-specific custom language model. Due to the default model not being fine-tuned for the medical domain, we resorted to building a custom model which focuses on recognizing medical entities.

Using Watson NLU involves sending HTTP requests to the service's API. Each request contains a chunk of text to be analyzed or a URL to a webpage, the textual content of which will

[13] https://www.ibm.com/watson/services/natural-language-understanding/
[14] https://www.ibm.com/cloud/

be retrieved and analyzed by the service, along with a set of configuration parameters. Similar to many other Watson services, Watson NLU also comes with SDKs for different programming languages such as Java, Node.js and Python. Since our prototype was developed in Node.js, we used Watson NLU via its Node.js SDK.

4.2.2 IBM Watson Knowledge Studio

In order to create a custom model for Watson NLU, we used *IBM Watson Knowledge Studio (WKS)*[15]. WKS is a stand-alone product aiming at better involving field-experts in the training of supervised machine learning language models to process unstructured data. The product offers a user-friendly interface and features which enable collaboration through iterative processes.

Key artifacts of a WKS project include: (1) a type system defining entity and relationship types; (2) three types of annotation components: a dictionary pre-annotator, a rule-based annotator and a machine learning annotator; and (3) documents to train and evaluate annotation components. Currently, only rule-based and machine learning annotation components created using WKS can be deployed as custom models of Watson NLU instances.

For the scope of this project, due to the limited resources and the unavailability of field-experts, we decided on constructing an extensive collection of dictionaries and deploy it via a rule-based annotator. These dictionaries can later be used to pre-annotate documents to assist human annotators with a set of preliminary annotations.

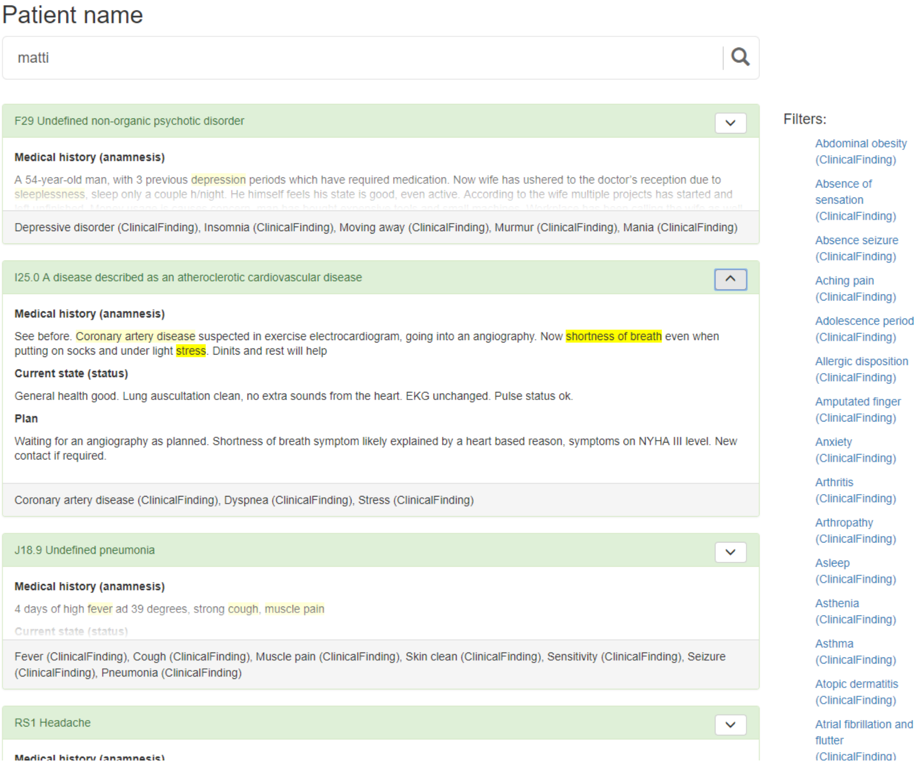
[15] https://www.ibm.com/watson/services/knowledge-studio/

4.3 The First Prototype

4.3.1 Use-case Overview

As previously mentioned, this prototype is that of a general DSS aiming at helping doctors to process the unstructured portion of notes from patient visits more efficiently and accurately. The DSS achieves this by highlighting keywords from notes and shows the relations between them and sections of the guideline document.

The prototype has a search function to search doctor notes using a patient’s name. For demonstration purposes, the database has only one patient called Matti with multiple doctor notes. After a search request is made, the results will be shown under the input area as result cards. Each of these cards consists of a header which tells the main diagnosis and a body that contains the patient’s medical history, current state and a suggested treatment plan. At the bottom of the card, the names of the detected entities and their types are shown as tags. Next to the search results is a list of all the distinct entities found in the base document which can be used as a filter to single out notes in which the selected entity appears. An example of search results can be seen in Figure 11.



Patient name

matti

F29 Undefined non-organic psychotic disorder

Medical history (anamnesis)

A 54-year-old man, with 3 previous depression periods which have required medication. Now wife has ushered to the doctor's reception due to sleeplessness, sleep only a couple h/night. He himself feels his state is good, even active. According to the wife multiple projects has started and left unfinished. Money usage is causes concern, man has bought expensive tools and small machines. Workplace has been calling the wife as well

Depressive disorder (ClinicalFinding), Insomnia (ClinicalFinding), Moving away (ClinicalFinding), Murmur (ClinicalFinding), Mania (ClinicalFinding)

I25.0 A disease described as an atheroclerotic cardiovascular disease

Medical history (anamnesis)

See before. Coronary artery disease suspected in exercise electrocardiogram, going into an angiography. Now shortness of breath even when putting on socks and under light stress. Dinits and rest will help

Current state (status)

General health good. Lung auscultation clean, no extra sounds from the heart. EKG unchanged. Pulse status ok.

Plan

Waiting for an angiography as planned. Shortness of breath symptom likely explained by a heart based reason, symptoms on NYHA III level. New contact if required.

Coronary artery disease (ClinicalFinding), Dyspnea (ClinicalFinding), Stress (ClinicalFinding)

J18.9 Undefined pneumonia

Medical history (anamnesis)

4 days of high fever ad 39 degrees, strong cough, muscle pain

Current state (status)

Fever (ClinicalFinding), Cough (ClinicalFinding), Muscle pain (ClinicalFinding), Skin clean (ClinicalFinding), Sensitivity (ClinicalFinding), Seizure (ClinicalFinding), Pneumonia (ClinicalFinding)

RS1 Headache

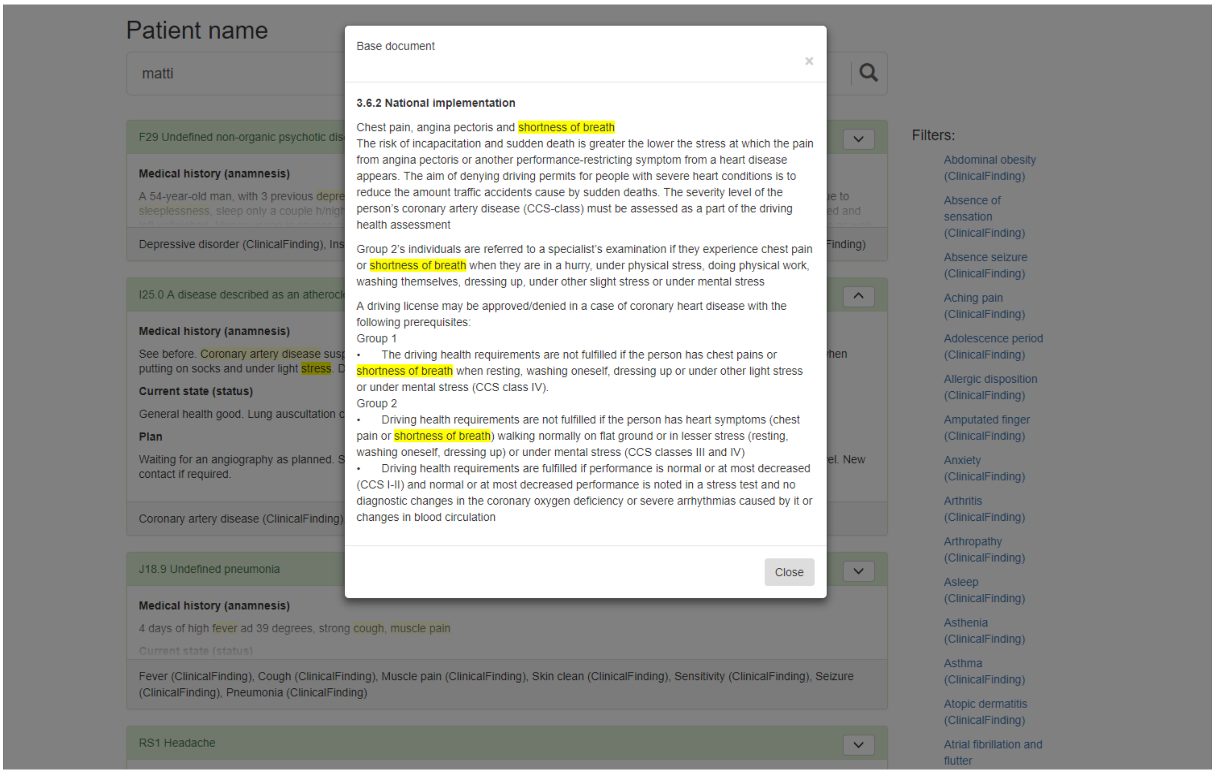
Medical history (anamnesis)

Filters:

- Abdominal obesity (ClinicalFinding)
- Absence of sensation (ClinicalFinding)
- Absence seizure (ClinicalFinding)
- Aching pain (ClinicalFinding)
- Adolescence period (ClinicalFinding)
- Allergic disposition (ClinicalFinding)
- Amputated finger (ClinicalFinding)
- Anxiety (ClinicalFinding)
- Arthritis (ClinicalFinding)
- Arthropathy (ClinicalFinding)
- Asleep (ClinicalFinding)
- Asthenia (ClinicalFinding)
- Asthma (ClinicalFinding)
- Atopic dermatitis (ClinicalFinding)
- Atrial fibrillation and flutter (ClinicalFinding)

Figure 11. First prototype – Example of search results

Parts of the text are highlighted if they are recognized as an entity by the system. There are two types of highlighting: texts highlighted in light grey indicate entities that do not have related base document excerpts and texts highlighted in bright yellow signify entities that do. Upon clicking a bright-yellow highlight, a modal dialog box containing related base document extracts – i.e. sections of the base documents with presences of the selected entity – will be opened to provide references. These extracts also have said entity highlighted, as shown in Figure 12.



Patient name
matti
F29 Undefined non-organic psychotic dis
I25.0 A disease described as an atherocl
J18.9 Undefined pneumonia
RS1 Headache
Filters:
Abdominal obesity (ClinicalFinding)
Absence of sensation (ClinicalFinding)
Absence seizure (ClinicalFinding)
Aching pain (ClinicalFinding)
Adolescence period (ClinicalFinding)
Allergic disposition (ClinicalFinding)
Amputated finger (ClinicalFinding)
Anxiety (ClinicalFinding)
Arthritis (ClinicalFinding)
Arthropathy (ClinicalFinding)
Asleep (ClinicalFinding)
Asthenia (ClinicalFinding)
Asthma (ClinicalFinding)
Atopic dermatitis (ClinicalFinding)
Atrial fibrillation and flutter
Close
×

Base document

3.6.2 National implementation

Chest pain, angina pectoris and shortness of breath
The risk of incapacitation and sudden death is greater the lower the stress at which the pain from angina pectoris or another performance-restricting symptom from a heart disease appears. The aim of denying driving permits for people with severe heart conditions is to reduce the amount traffic accidents cause by sudden deaths. The severity level of the person's coronary artery disease (CCS-class) must be assessed as a part of the driving health assessment

Group 2's individuals are referred to a specialist's examination if they experience chest pain or shortness of breath when they are in a hurry, under physical stress, doing physical work, washing themselves, dressing up, under other slight stress or under mental stress

A driving license may be approved/denied in a case of coronary heart disease with the following prerequisites:
Group 1

- The driving health requirements are not fulfilled if the person has chest pains or shortness of breath when resting, washing oneself, dressing up or under other light stress or under mental stress (CCS class IV).

Group 2

- Driving health requirements are not fulfilled if the person has heart symptoms (chest pain or shortness of breath) walking normally on flat ground or in lesser stress (resting, washing oneself, dressing up) or under mental stress (CCS classes III and IV)
- Driving health requirements are fulfilled if performance is normal or at most decreased (CCS I-II) and normal or at most decreased performance is noted in a stress test and no diagnostic changes in the coronary oxygen deficiency or severe arrhythmias caused by it or changes in blood circulation

Medical history (anamnesis)

A 54-year-old man, with 3 previous depre
sleeplessness, sleep only a couple h/nigh

Depressive disorder (ClinicalFinding), Ins

Medical history (anamnesis)

See before. Coronary artery disease susp
putting on socks and under light stress. D

Current state (status)

General health good. Lung auscultation c

Plan

Waiting for an angiography as planned. S
contact if required.

Coronary artery disease (ClinicalFinding)

Medical history (anamnesis)

4 days of high fever ad 39 degrees, strong cough, muscle pain

Current state (status)

Fever (ClinicalFinding), Cough (ClinicalFinding), Muscle pain (ClinicalFinding), Skin clean (ClinicalFinding), Sensitivity (ClinicalFinding), Seizure (ClinicalFinding), Pneumonia (ClinicalFinding)

Figure 12. First prototype – Example of reference modal box

4.3.2 Prototype Architecture

As our formulation of TODSS has been an incremental process, its first protype did not necessarily conform to the outlined general architecture. The most obvious deviation is the lack of the *Relevance Ranking* module due to the ontology being used simply as a dictionary at this point (see Figure 13).

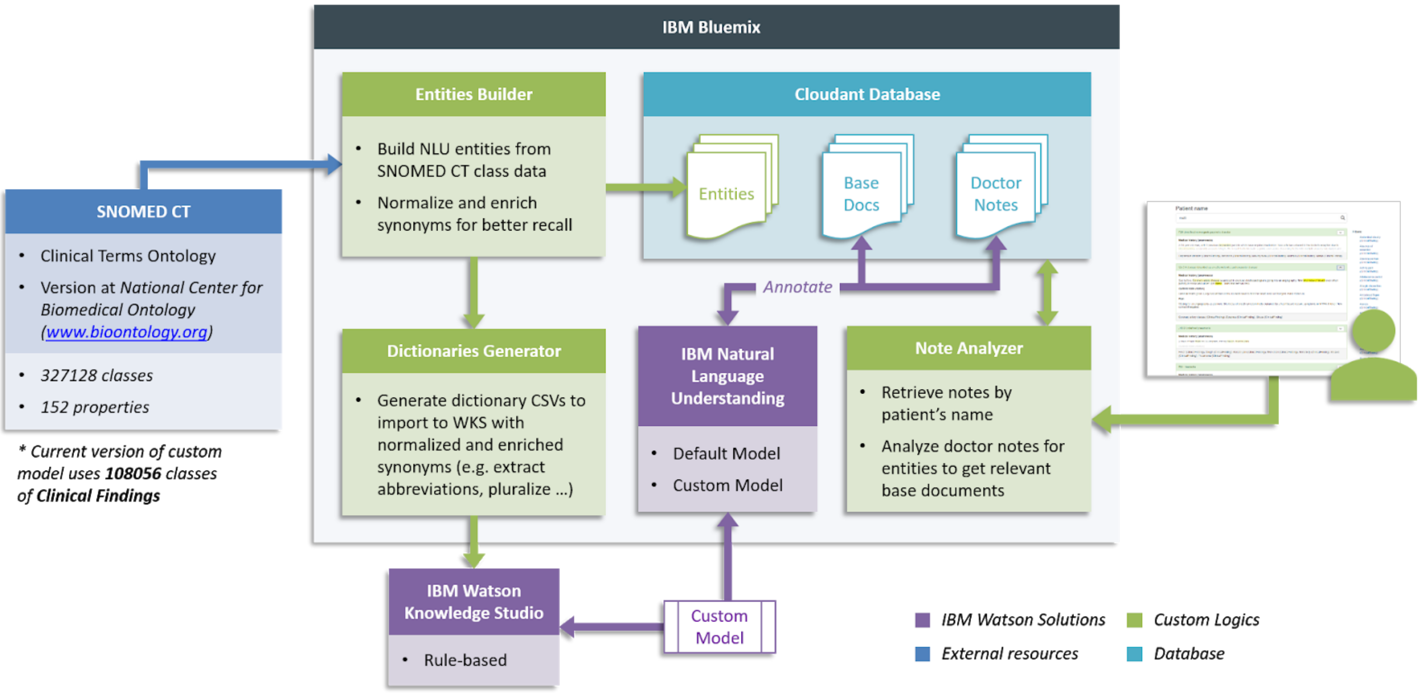

IBM Bluemix
Entities Builder
Build NLU entities from SNOMED CT class data
Normalize and enrich synonyms for better recall
Cloudant Database
Entities
Base Docs
Doctor Notes
SNOMED CT
Clinical Terms Ontology
Version at National Center for Biomedical Ontology (www.bioontology.org)
327128 classes
152 properties
* Current version of custom model uses 108056 classes of Clinical Findings
Annotate
Dictionaries Generator
Generate dictionary CSVs to import to WKS with normalized and enriched synonyms (e.g. extract abbreviations, pluralize ...)
IBM Natural Language Understanding
Default Model
Custom Model
Note Analyzer
Retrieve notes by patient's name
Analyze doctor notes for entities to get relevant base documents
Patient name
IBM Watson Knowledge Studio
Rule-based
Custom Model
IBM Watson Solutions
Custom Logics
External resources
Database

Figure 13. Architecture of the first TODSS prototype

4.3.2.1 Entities Builder

In our first prototype, the *Entities Builder* module assumes the role of the Ontology Interfacing module in the general architecture. In addition to retrieving SNOMED-CT ontology data from BioPortal, the module was made to achieve two additional objectives: prepare and enrich data for Dictionaries Generator; and construct an entity database for disambiguation.

Each class of the SNOMED-CT ontology comes with a collection of synonyms, which we used to generate surface forms of entities. However, these synonyms are not always optimized for the purpose of text extraction. Through our Entity Builder, we normalized the synonyms, removed redundancies and stop words, extracted abbreviations and applied pluralization/singularization for better recall. One thing to note is that currently, WKS dictionary entries match only texts in higher cases, therefore all surface forms except abbreviations should be in lower-case.

Entity analyses done with Watson NLU default model can provide disambiguation when applicable. In those cases, a link to a DbPedia resource page is given to specifically identify an entity (e.g. *http://dbpedia.org/resource/CNN*). This is not the case for analyses done with custom models. In order to overcome this limitation, we created our own entity database for disambiguation. Each entity is represented by a database entry, which has an _id, which in turn is the sub-path of the entity's URI on BioPortal (e.g. "410006001" - *http://purl.bioontology.org/ontology/SNOMEDCT/410006001* - DRE), a name which is the Preferred Name provided by the ontology, a list of surface forms, a main type and a list of subtypes.

4.3.2.2 Dictionary Generator

Each WKS project can make use of up to *64* dictionaries of *15000* entries each. A dictionary entry consists of a lemma, a list of surface forms and a designated part of speech. The dictionaries can either be created manually or transferred from another WKS project (the entire dictionary collection is exported and imported as a .zip file). Likewise, entries within a dictionary can be added individually or imported from .csv files (maximum 1 MB per file).

To meet the constraints set by WKS, we implemented a module to generate .csv files to be used in the creations of dictionaries. Since WKS employs certain non-disclosed methods to ensure only dictionaries exported from another WKS can be imported in bulk we had to create the dictionaries manually and import the entries using .csv files. It would be a lot more convenient if WKS could expose some APIs for these functionalities.

4.3.2.3 Document Databases and Document Handler

In a real-world scenario, there would exist databases for doctor notes and base documents. These documents would be annotated as soon as they are added to the DSS. In this prototype, we implemented a *Document Handler* module to simulate this real-world scenario. Input documents were imported to the module in two text files, along with two corresponding *.json* files specifying the template of the doctor notes and the structure of the base document which guides the segmentation process. The segmented texts were treated as separate documents. These documents were then analyzed by Watson NLU for entity mentions. Based on the detected surfaces, disambiguation data were retrieved from the entity database and in turn used to annotate the documents. Annotated document objects were then saved to the databases, with each object containing the document's text and a list of entities found within said text and their positions within the text.

The Document Handler module, together with the Entity Builder module, Watson NLU and WKS, made up the *NLP* component of the prototype

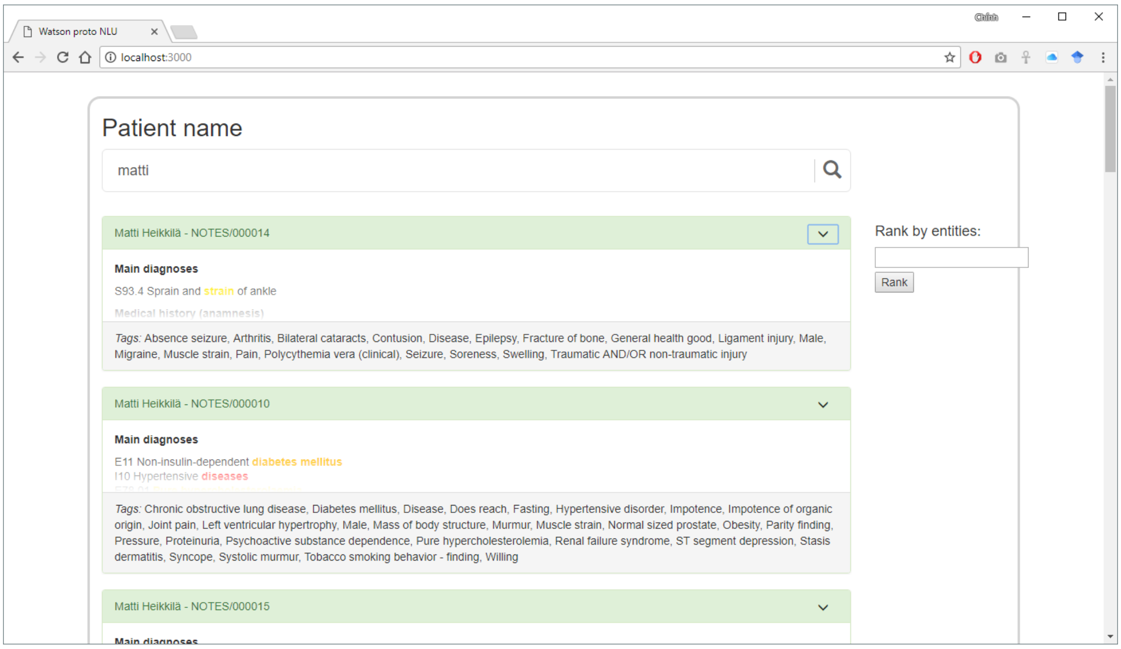
4.3.2.4 Note Analyzer

The back-end logics of the server component of the first prototype is contained in the *Note Analyzer* module. Using the patient name inputted via the web-based front-end, this module retrieves doctor notes from the patient's medical records. Entities found in the doctor notes are then used to retrieve relevant parts of the base document - parts in which there are mentions of such entities. A front-end-friendly response is then generated by the module, in which entities are mapped to the base document's parts if applicable and marked for highlighting.

4.4 The Second Prototype

4.4.1 Use-case Overview

Our main goal for the second prototype was to include semantic similarity measurements into the DSS and as a result, further realizing the described general use-cases. The user-story of this prototype does not differ greatly from the first. Introduced to each step of the story, however, is a considerable amount of augmented user insights through a variety of semantic-based ranking operations carried out by the DSS. This is best illustrated by the following updated screen-cap demonstration.



Patient name

matti

Matti Heikkilä - NOTES/000014

Main diagnoses

S93.4 Sprain and strain of ankle

Medical history (anamnesis)

Tags: Absence seizure, Arthritis, Bilateral cataracts, Contusion, Disease, Epilepsy, Fracture of bone, General health good, Ligament injury, Male, Migraine, Muscle strain, Pain, Polycythemia vera (clinical), Seizure, Soreness, Swelling, Traumatic AND/OR non-traumatic injury

Matti Heikkilä - NOTES/000010

Main diagnoses

E11 Non-insulin-dependent diabetes mellitus
I10 Hypertensive diseases

Tags: Chronic obstructive lung disease, Diabetes mellitus, Disease, Does reach, Fasting, Hypertensive disorder, Impotence, Impotence of organic origin, Joint pain, Left ventricular hypertrophy, Male, Mass of body structure, Murmur, Muscle strain, Normal sized prostate, Obesity, Parity finding, Pressure, Proteinuria, Psychoactive substance dependence, Pure hypercholesterolemia, Renal failure syndrome, ST segment depression, Stasis dermatitis, Syncope, Systolic murmur, Tobacco smoking behavior - finding, Willing

Matti Heikkilä - NOTES/000015

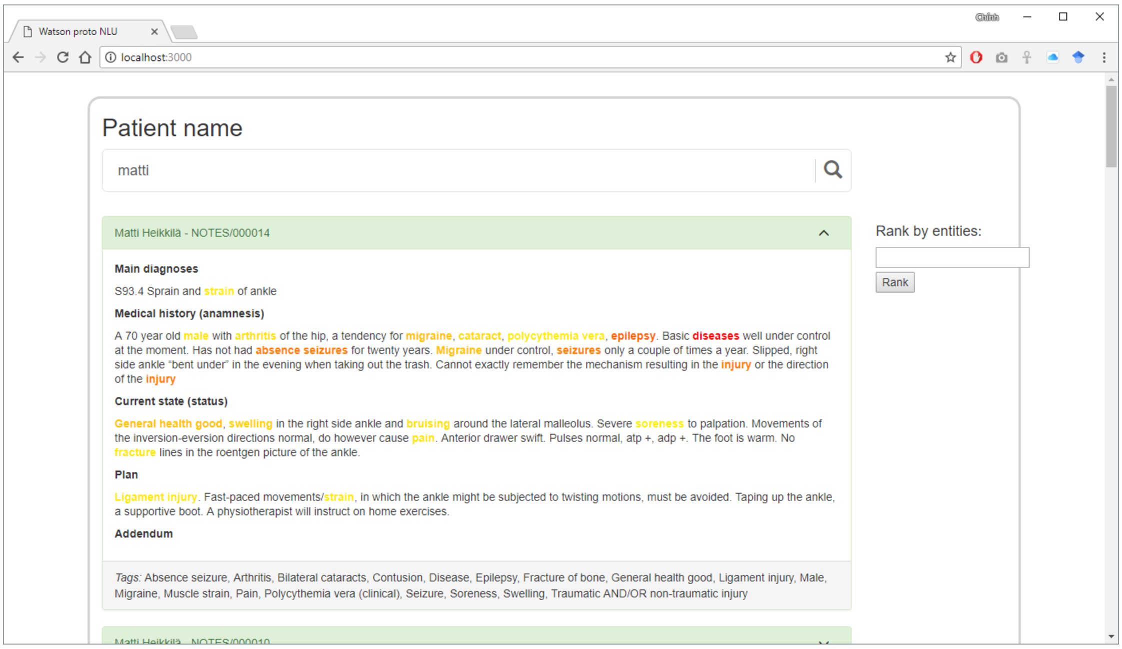
Main diagnoses

Rank by entities:

Rank

Figure 14. Second prototype – Ranked search results

Instead of only retrieving patient records, the DSS now ranks the records by the degree of relevancy of each to the guideline document. Doctor notes in which detected medical terms are considered more relevant to the regulations in the guideline document are presented first, as shown in Figure 14. In this way, the most important and relevant cues are prioritized for display over the less significant or irrelevant ones, which tends to make up the majority of the patient's data. As the number of patient's records grows, the benefits of a ranking mechanism are amplified.



Matti Heikkilä - NOTES/000014

Main diagnoses

S93.4 Sprain and strain of ankle

Medical history (anamnesis)

A 70 year old male with arthritis of the hip, a tendency for migraine, cataract, polycythemia vera, epilepsy. Basic diseases well under control at the moment. Has not had absence seizures for twenty years. Migraine under control, seizures only a couple of times a year. Slipped, right side ankle "bent under" in the evening when taking out the trash. Cannot exactly remember the mechanism resulting in the injury or the direction of the injury

Current state (status)

General health good, swelling in the right side ankle and bruising around the lateral malleolus. Severe soreness to palpation. Movements of the inversion-eversion directions normal, do however cause pain. Anterior drawer swift. Pulses normal, atp +, adp +. The foot is warm. No fracture lines in the roentgen picture of the ankle.

Plan

Ligament injury. Fast-paced movements/strain, in which the ankle might be subjected to twisting motions, must be avoided. Taping up the ankle, a supportive boot. A physiotherapist will instruct on home exercises.

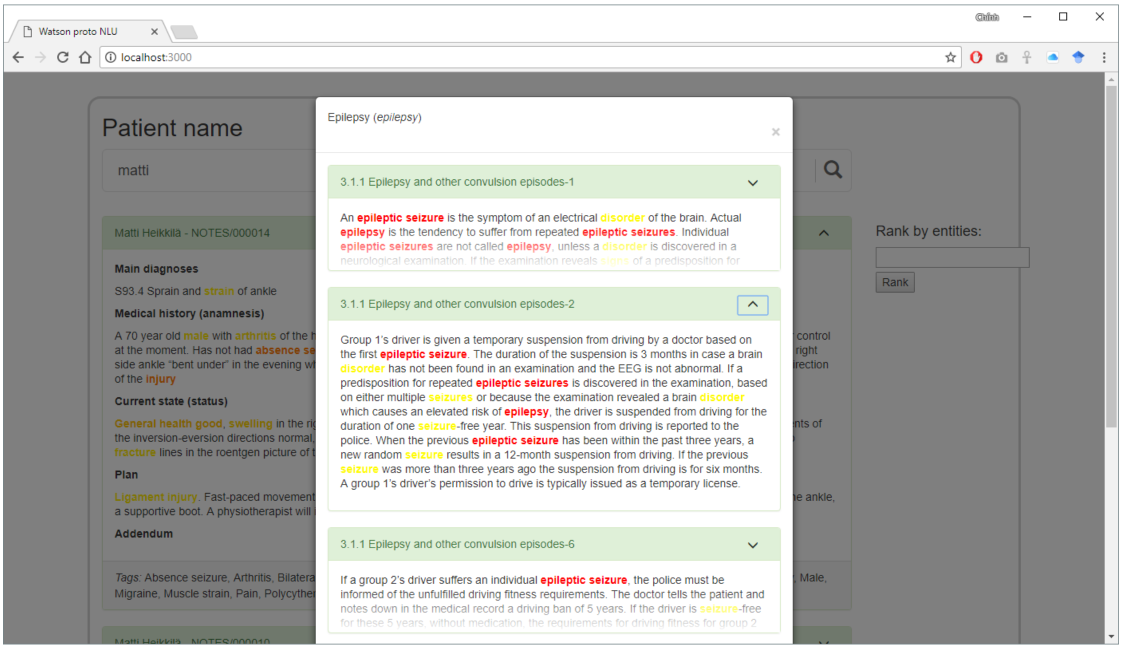
Addendum

Tags: Absence seizure, Arthritis, Bilateral cataracts, Contusion, Disease, Epilepsy, Fracture of bone, General health good, Ligament injury, Male, Migraine, Muscle strain, Pain, Polycythemia vera (clinical), Seizure, Soreness, Swelling, Traumatic AND/OR non-traumatic injury

Matti Heikkilä - NOTES/000010

Figure 15. Second prototype - Color-coded entities in patient records

When the user clicks on a record to expand it for further inspection, words representing medical concepts can be seen to be highlighted with different color intensities. These differences proportionally reflect their degree of relevancy to the entire guideline document with bright-red-colored terms being the most and light-yellow ones being the least relevant. This feature helps the user to quickly identify risk factors in each note.


Watson proto NLU
localhost:3000
Patient name
matti
Matti Heikkilä - NOTES/000014
Main diagnoses
S93.4 Sprain and strain of ankle
Medical history (anamnesis)
A 70 year old male with arthritis of the h
at the moment. Has not had absence se
side ankle "bent under" in the evening wh
of the injury
Current state (status)
General health good, swelling in the rig
the inversion-eversion directions normal,
fracture lines in the roentgen picture of t
Plan
Ligament injury. Fast-paced movement
a supportive boot. A physiotherapist will i
Addendum
Tags: Absence seizure, Arthritis, Bilatera
Migraine, Muscle strain, Pain, Polycyther
Matti Heikkilä - NOTES/000010
Rank by entities:
Rank

Epilepsy (*epilepsy*)

3.1.1 Epilepsy and other convulsion episodes-1

An **epileptic seizure** is the symptom of an electrical **disorder** of the brain. Actual **epilepsy** is the tendency to suffer from repeated **epileptic seizures**. Individual **epileptic seizures** are not called **epilepsy**, unless a **disorder** is discovered in a neurological examination. If the examination reveals **signs** of a predisposition for

3.1.1 Epilepsy and other convulsion episodes-2

Group 1's driver is given a temporary suspension from driving by a doctor based on the first **epileptic seizure**. The duration of the suspension is 3 months in case a brain **disorder** has not been found in an examination and the EEG is not abnormal. If a predisposition for repeated **epileptic seizures** is discovered in the examination, based on either multiple **seizures** or because the examination revealed a brain **disorder** which causes an elevated risk of **epilepsy**, the driver is suspended from driving for the duration of one **seizure**-free year. This suspension from driving is reported to the police. When the previous **epileptic seizure** has been within the past three years, a new random **seizure** results in a 12-month suspension from driving. If the previous **seizure** was more than three years ago the suspension from driving is for six months. A group 1's driver's permission to drive is typically issued as a temporary license.

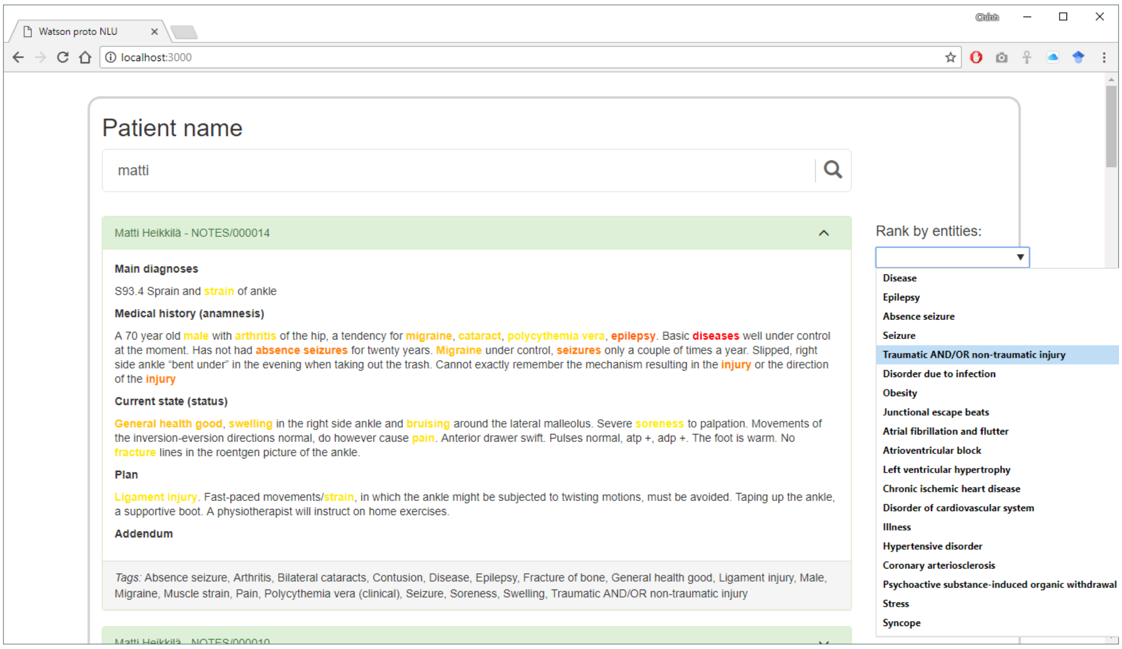
3.1.1 Epilepsy and other convulsion episodes-6

If a group 2's driver suffers an individual **epileptic seizure**, the police must be informed of the unfulfilled driving fitness requirements. The doctor tells the patient and notes down in the medical record a driving ban of 5 years. If the driver is **seizure**-free for these 5 years, without medication, the requirements for driving fitness for group 2

Figure 16. Second prototype – Reference modal box of ranked

The contents displayed in the reference modal box are also enhanced. While in the first prototype, only extracts of the guideline document with exact mentions of the selected concept were provided, in this prototype, the DSS selects the top five most relevant extracts with respect to the concept in question. Not only does this approach not miss out sections with exact mentions due to high bias toward exact matches (semantic similarity score is *1*), it also includes sections with high topical relevance to the selected concept that do not contain any of its surface forms, thus noticeably improving recalls.

Entities which are present in these extracts are also color-coded similar to those found in doctor notes, only in this case, the color intensities signify their similarity to the selected term (see Figure 16).


Watson proto NLU
localhost:3000
Rank by entities:
Disease
Epilepsy
Absence seizure
Seizure
Traumatic AND/OR non-traumatic injury
Disorder due to infection
Obesity
Junctional escape beats
Atrial fibrillation and flutter
Atrioventricular block
Left ventricular hypertrophy
Chronic ischemic heart disease
Disorder of cardiovascular system
Illness
Hypertensive disorder
Coronary arteriosclerosis
Psychoactive substance-induced organic withdrawal
Stress
Syncope

Patient name

matti

Matti Heikkilä - NOTES/000014

Main diagnoses

S93.4 Sprain and strain of ankle

Medical history (anamnesis)

A 70 year old male with arthritis of the hip, a tendency for **migraine**, **cataract**, **polycythemia vera**, **epilepsy**. Basic **diseases** well under control at the moment. Has not had **absence seizures** for twenty years. **Migraine** under control, **seizures** only a couple of times a year. Slipped, right side ankle "bent under" in the evening when taking out the trash. Cannot exactly remember the mechanism resulting in the **injury** or the direction of the **injury**

Current state (status)

General health good, **swelling** in the right side ankle and bruising around the lateral malleolus. Severe soreness to palpation. Movements of the inversion-eversion directions normal, do however cause pain. Anterior drawer swift. Pulses normal, atp +, adp +. The foot is warm. No fracture lines in the roentgen picture of the ankle.

Plan

Ligament injury. Fast-paced movements/strain, in which the ankle might be subjected to twisting motions, must be avoided. Taping up the ankle, a supportive boot. A physiotherapist will instruct on home exercises.

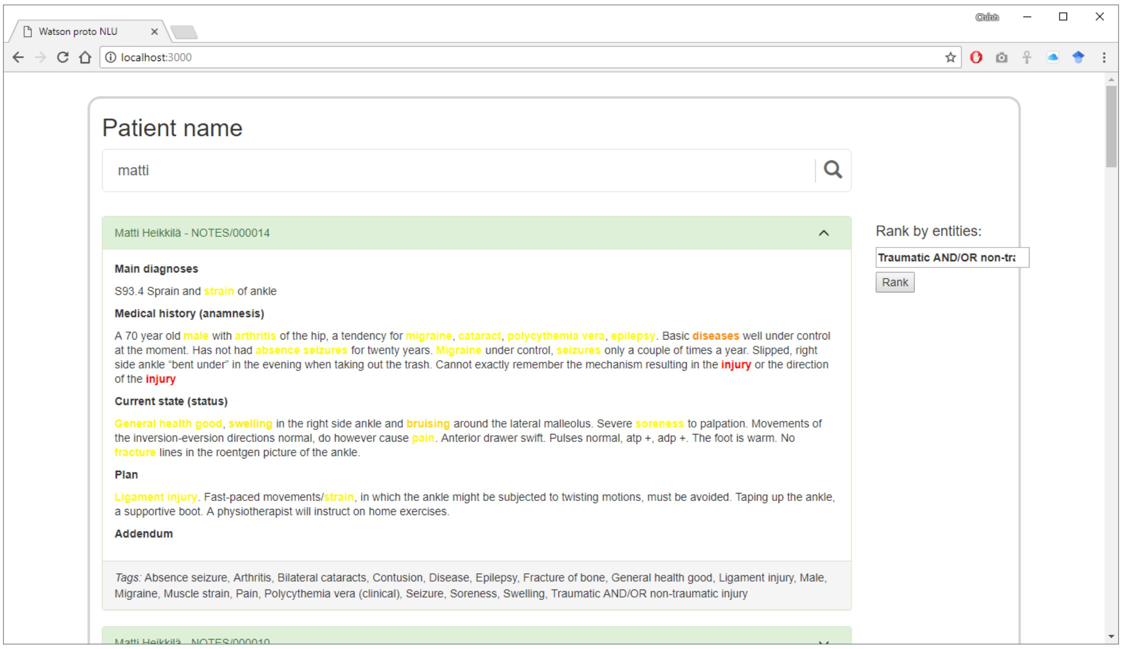
Addendum

Tags: Absence seizure, Arthritis, Bilateral cataracts, Contusion, Disease, Epilepsy, Fracture of bone, General health good, Ligament injury, Male, Migraine, Muscle strain, Pain, Polycythemia vera (clinical), Seizure, Soreness, Swelling, Traumatic AND/OR non-traumatic injury

Matti Heikkilä - NOTES/000010

Figure 17. Second prototype – Rank-by-entities drop-down menu

Alternatively, the user can choose to rank the patient records relative to a specific concept (Figure 17). In the drop-down menu replacing the filter list in the first prototype, a list of concepts ordered by their relevance to the whole guideline document can be found. Upon the user selecting a concept from this list, all doctor notes are rearranged in descending order of semantic relevance scores between them and said concept, as shown in Figure 18. The colors of their contained terms are also updated accordingly.



Matti Heikkilä - NOTES/000014

Main diagnoses

S93.4 Sprain and strain of ankle

Medical history (anamnesis)

A 70 year old male with arthritis of the hip, a tendency for migraine, cataract, polycythemia vera, epilepsy. Basic diseases well under control at the moment. Has not had absence seizures for twenty years. Migraine under control, seizures only a couple of times a year. Slipped, right side ankle "bent under" in the evening when taking out the trash. Cannot exactly remember the mechanism resulting in the injury or the direction of the injury

Current state (status)

General health good, swelling in the right side ankle and bruising around the lateral malleolus. Severe soreness to palpation. Movements of the inversion-eversion directions normal, do however cause pain. Anterior drawer swift. Pulses normal, atp +, adp +. The foot is warm. No fracture lines in the roentgen picture of the ankle.

Plan

Ligament injury. Fast-paced movements/strain, in which the ankle might be subjected to twisting motions, must be avoided. Taping up the ankle, a supportive boot. A physiotherapist will instruct on home exercises.

Addendum

Tags: Absence seizure, Arthritis, Bilateral cataracts, Contusion, Disease, Epilepsy, Fracture of bone, General health good, Ligament injury, Male, Migraine, Muscle strain, Pain, Polycythemia vera (clinical), Seizure, Soreness, Swelling, Traumatic AND/OR non-traumatic injury

Matti Heikkilä - NOTES/000010

Figure 18. Second prototype – Rank-by-entities results

4.4.2 Prototype Architecture

As an immediate step-up from the first prototype, our second prototype of TODSS retains many characteristic components and features of its predecessor. However, analyses based on semantic similarities and relevance assessments have now been introduced into the DSS. This was achieved through the implementation of semantic-enabled entity databases and two new logic modules: *Relevance* Scorer and *Ranker/Analyzer*. The resemblance to the general architecture can be seen much clearer in this prototype, as illustrated in Figure 19.

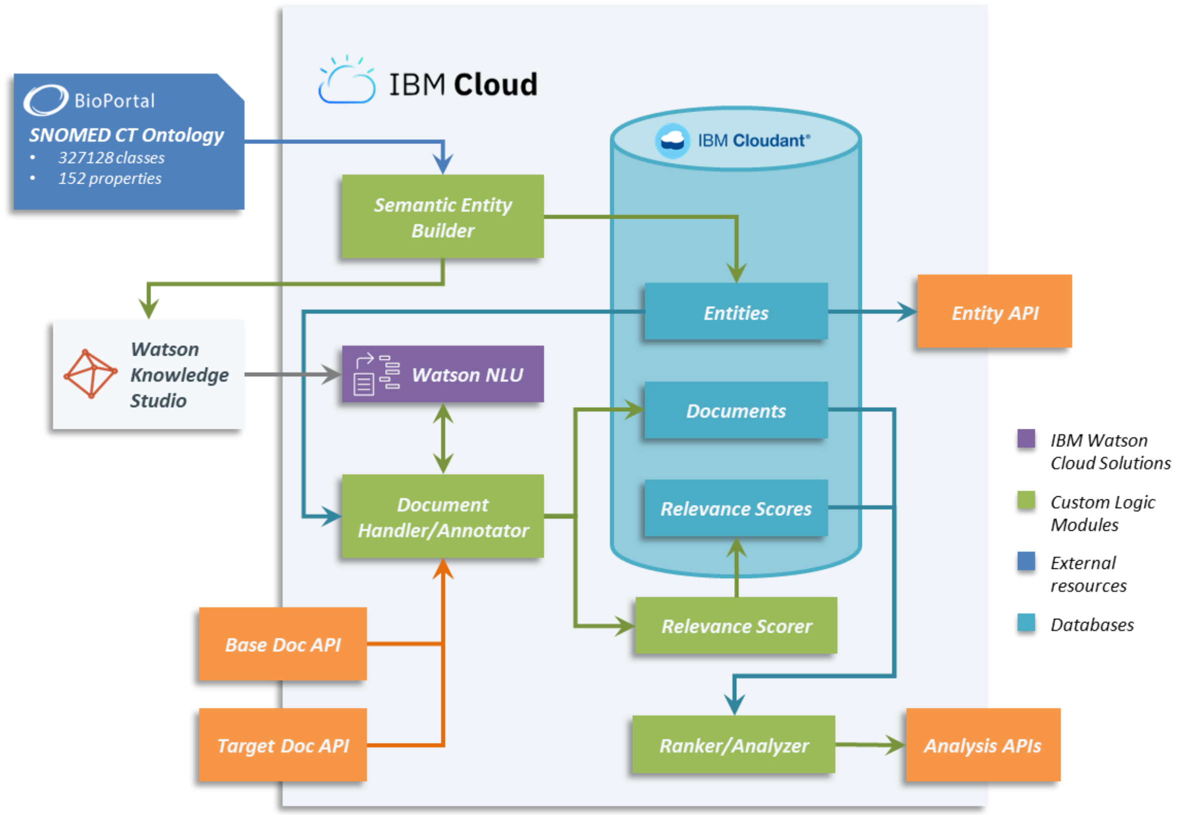


Figure 19. Architecture of the second TODSS prototype

4.4.2.1 Semantic Entity Builder and Entity Databases

In the first prototype, we only retrieved and stored descriptive data of SNOMED-CT classes not their taxonomical information. This decision, partly due to SNOMED-CT restriction on storing term IDs without proper licenses, barred us from utilizing the full informational potential of the ontology. To circumvent this, we first implemented a sub-module to the existing Entity Builder module named *Semantic Enabler*, in which we define aspects of the DSS as metadata, including the context of our own ontology of medical concepts. We then aligned the SCOMED-CT ontology to ours by mapping relevant properties of its classes to our classes within said context.

As the second step, we modified the existing non-SQL databases for ontology-based entities to effectively store structural information of classes. Each entry in these databases was modified to have its data properties complying with *JSON-LD* standards and appended with the expanded list of URIs of the super-classes of the entity it represented. While having all superclass URIs stored for each entity increased the size of the database to some extent, it eliminated the need of traversing the graph each time a concept-level semantic similarity score was to be calculated, thus significantly improving the performance of the DSS.

4.4.2.2 Relevance Scorer

The *Relevance Scorer* is yet another improvement in our second prototype of TODSS. Serving as a utility module, it houses the logics for semantic similarity calculations and is invoked by the Document Handler module as a new document comes into the DSS. Following the segmentation and entity annotation processes, relevance scores between the document and each of its entities are calculated and embedded in the document database entry. After that, depending on the type of the received document, the module computes the semantic similarity scores between the document and all other existing documents of the opposite type. The results are also saved to a database, including all preliminary calculation results. This behavior adheres to the same logic behind the storing of the flattened subsumer branch of each concept: computationally expensive operations should only be carried out a minimum number of times.

4.4.2.3 Ranker/Analyzer

Ranking-based analyses – a new feature of this prototype - are performed by the reworked *Ranker/Analyzer* module. Using pre-calculated similarity and relevance scores, the module provides the frontend with five functionalities via HTTP APIs: (1) rank individual target documents by relevance with respects to all base documents, (2) rank concepts found in target documents by their relevance to all base documents, (3) rank base documents by their relevance to a selected concept, (4) rank target documents by their relevance to a selected concept and (5) rank concepts by their similarity to a selected concept. These functionalities fulfill all user needs outlined in the prototype use-case.

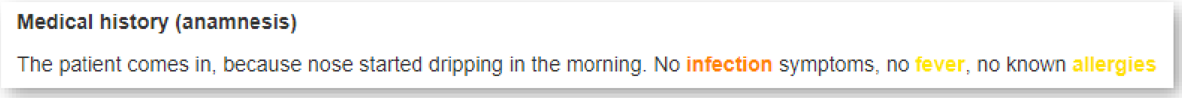
5 Evaluation of Proof-of-Concept Prototypes

The first of the two prototypes was demonstrated in several meetings and workshops within the scope of the Value from Public Health Data with Cognitive Computing project and received feedback from experts in various fields and other stakeholders. The consensus was that the prototype has succeeded in addressing a significant problem as well as proving the maturity of the utilized technologies and the practicability of the approach. It was also showed to have an advantage in comparison with other ready-made solutions in (Khriyenko et al., 2018).

While the second prototype is a clear improvement over the first, we did not have the same opportunities for evaluation since the development stretched beyond of the project's timeframe. However, in internal demonstrations, the increased utility of the prototype, thanks to the added ranking features and the validity of estimating relevance by semantic similarity measurements, was recognized by our peers.

Due to several constrains, mainly time and human resources, and despite promising initial results of the prototypes in proving the feasibility of a TODSS, there were limitations which we could not overcome to conclusively evaluate the effectiveness of our approach.

One major flaw we noticed during the development of the prototypes was the fact that the NLP modules could not distinguish between the positivity and the negativity of a term mention. This happened on some patient records where the doctor made active statements about the non-presence (negation) of some symptoms. As an example, in the following extract of one note (Figure 20), the doctor wrote:



Medical history (anamnesis)

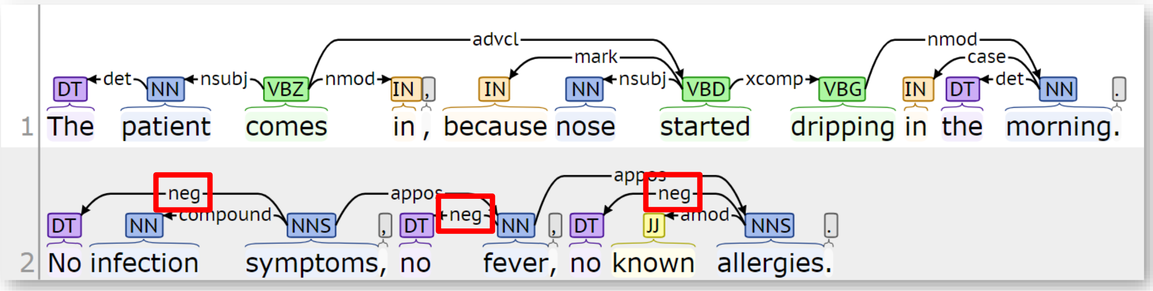
The patient comes in, because nose started dripping in the morning. No **infection** symptoms, no **fever**, no known **allergies**

Figure 20. Example of missed negations

"The patient comes in, because nose started dripping in the morning. No infection symptoms, no fever, no known allergies."

While the intention of the writer in this case was to explicitly deny the presences of certain terms, the DSS prototype took them as positive hits, and counted them as such toward future calculations. Cases similar to this also occurred with the guideline document.

Since our current NLP solution, IBM Watson NLU, lacks the capability to provide negation detection for entities, we have been considering and, in the future prototypes, will utilize other NLP solutions with one notable option being *Stanford CoreNLP* – a novel NLP toolkit first proposed in (Manning et al., 2014). The same text analyzed using Stanford CoreNLP online demo[16], yielded the following result (Figure 21):


1 The patient comes in , because nose started dripping in the morning.

2 No infection symptoms, no fever, no known allergies.

Figure 21. Example of NLP analyses performed by Stanford CoreNLP

As shown in this example, the negation relations are reliably detected by the software through the positioning of determiners *"no"* preceding the nouns. This result is significantly better than what we had with Watson NLU.

Unfortunately, however, we have not been able to substitute Watson NLU with Stanford CoreNLP in our final prototype since more effort need to be taken to enable the software to work with customized vocabulary sets so that ontology-based entities can be integrated in its analysis. Nevertheless, we learned of the inaccuracies caused by ignored negations and will make the correction of this issue a main focus in future prototypes.

Another limitation we faced was in evaluating the usefulness of the prototypes. We attempted to organize a session in which a doctor – a supposed real-life user of the DSS – was to give feedback on its performance. However, due to the busy nature of his work, the doctor

[16] http://nlp.stanford.edu:8080/corenlp/process

was not able to allocate us any time for such a session during the Value from Public Health Data with Cognitive Computing project. As the project phased out, this was no longer considered a priority. We perfectly understand that the lack of a formal and un-biased evaluation process with the direct involvement of domain experts makes any claim of validity and effectiveness of a DSS questionable. Therefore, in the continuation of the prototype development, we will emphasize the implementing a comprehensive testing framework.

As for potential additional features that we would like to explore and incorporate into the DSS, we prioritize adapting the analyses to use lemmatized representations of texts as a path to scale the solution to other languages. In language study, *lemmas* are canonical forms of words and usually used as dictionary entries – thus otherwise known as dictionary forms. For example, the sentence

Karjalanpiirakka on perinteinen suomalainen leivonnainen, jossa ohuen hapattamattoman ruiskuoren sisällä on riisipuuroa, ohraryynipuuroa tai perunasosetta.

in Finnish has the lemmatized form of

karjalanpiirak on perintein suomalain leivonnain, jos ohue hapattamattom ruiskuor sisä on riisipuuro , ohraryynipuuro tai perunasos.

For languages where word forms are complicated with agglutinations such as German or Finnish, lemmas provide a much more consistent way of denoting a sentence. As a result, entity annotation carried out on lemmatized text normally achieves improved performance. We have planned to implement another prototype which can experiment directly on the original un-translated Finnish documents in the near future.

The latest version of our prototype is open for evaluation by following instructions outlined in Appendix A.

6 Related Works

Watson Discovery[17] (previously known as ***Watson Retrieve and Rank***), being a general-purpose text-based solution, is moderately similar to our design of a TODSS. However, all analyses performed by Watson Discovery as the basis for its features are based only on the distribution of words within the managed corpus. More specifically, its underlying algorithms are that of *Apache Solr*[18] search engine, upon which the solution is built. In other words, it lacks the ability to specify a context in which document contents are understood. This results in inferior performance compared to our proposed approach, as reported in (Khriyenko et al., 2018).

As for medical domain, ***WatsonPaths*** (Lally et al., 2017), a scenario-based natural language question answering system, claims to be able to generate suggestions for diagnoses and treatment plans by applying a probabilistic inference graph starting with the description of the patient's condition and through a set of annotated and indexed facts extracted from medical literatures. By description, this solution better resembles the knowledge-driven DSS than a text-based system. Thus, the solution is vulnerable to its previously discussed limitations. A comprehensive comparison between our approach and WatsonPath was not possible, however, due to the lack of access.

Non-commercial related approaches to our work include the NLP-based EMR analysis applications presented in (Meystre & Haug, 2006) and (Byrd, Steinhubl, Sun, Ebadollahi, & Stewart, 2014). The first application, though maintaining a vocabulary set of health problems, did not organize this resource into an ontology from which structural information can be utilized. This design decision limited the utility of the application to only identifying health problems in medical records. The second application employed a more sophisticated hybrid NPL pipeline of both machine-learning and rule-based techniques which results in

[17] https://www.ibm.com/watson/services/discovery/
[18] http://lucene.apache.org/solr/

very high precision and recall. However, the focus of the application was restricted to detecting only signs of health failure and could not be easily adapted to other uses. The application also did not make use of any ontology-based analysis.

7 Conclusion

In this thesis, we examined the prospects of a text-based design to decision support systems in circumventing the challenges faced by traditional approaches in the Big Data era and how semantic technologies can be utilized to improve the performance of such systems. From the result of this theoretical analysis, we proposed the general design of an ontology-driven text-based decision support system (TODSS), of which two proof-of-concept prototypes were iteratively developed and evaluated.

Our evaluations in Chapter 2 showed that, at least in theory, a text-based DSS would have considerable advantages over traditional DSS-es while being able to alleviate the impacts of the Big Data challenge thanks to its ability to meaningfully utilize large amounts of textual data to support the decision-making process without the need of a centralized and complex knowledge-base or data model. This characteristic also enables a DSS of this type to be domain-independent, which result in high scalability as demonstrated by use-case examples in Section 2.2.2.

Ontologies and the Semantic Web, two of the hallmarks of semantic technologies were also proven to be well-suited for the extraction and representation of knowledge in textual artifacts within the context of a text-based DSS. A more in-depth analysis showed that by better utilizing the embedded structural information of ontologies, the information retrieval function of the DSS can be considerably improved. More specifically, the application of the concept-to-document and document-to-document semantic similarity measures proposed in Section 3.3.2, documents can be retrieved by their degrees of topical relevance without the need of probabilistic models, which require pre-analyses on large-scale corpuses. The presented semantic similarity measures were also shown to reflect the aspect of user-relevance in a text-based DSS, thus settled the ongoing *similarity-relevance* debate within this scope and serve as a valid answer to our second research question: ***“How to formalize and implement the concept of “relevance” in data filtering with-in the context of a text-based decision support?”***.

The feasibility and potential added values of a TODSS were effectively demonstrated by the implementation of our two prototypes of such a system. The first prototype was shown to outperform similar existing solutions (Khriyenko et al., 2018) and well-regarded by technical and medical experts in demonstrations during the Value from Public Health Data with Cognitive Computing project. As an incremental development from the previous, the second prototype offered a significantly increased recall, a more informative display of data and additional analysis options. These results convincingly proved the utilization of semantic similarity measures to be a viable option for improving the performance of a text-based DSS, therefore answering the third research question of our thesis: ***"How to boost the performance of a text-based decision-support system using semantic technologies?"***

Finally, despite the shortcomings and rooms for improvement as discussed in detail in Section **Error! Reference source not found.**, through theoretical arguments and practical evidences, we believe that ***a text-based ontology-driven approach is a robust and worthy of consideration option to designing an effective decision support system capable of rising above the challenge of unstructured big-data.***

Bibliography

References

Al-Mubaid, H., & Nguyen, H. A. (2006). A cluster-based approach for semantic similarity in the biomedical domain. Paper presented at the *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE,* 2713-2717.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web* (pp. 722-735) Springer.

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics, 44*(1), 118-125. 10.1016/j.jbi.2010.09.002 Retrieved from http://www.sciencedirect.com/science/article/pii/S1532046410001346

Bechhofer, S. (2009). OWL: Web ontology language. *Encyclopedia of database systems* (pp. 2008-2009) Springer.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American, 284*(5), 34-43.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research, 32*(suppl_1), D270.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the Association for Information Science and Technology, 54*(10), 913-925.

Byrd, R. J., Steinhubl, S. R., Sun, J., Ebadollahi, S., & Stewart, W. F. (2014). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International Journal of Medical Informatics, 83*(12), 983-992.

Choi, I., & Kim, M. (2003). Topic distillation using hierarchy concept tree. Paper presented at the *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval,* 371-372.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370-383. 10.1109/TKDE.2007.48 Retrieved from https://ieeexplore.ieee.org/document/4072748

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*(1), 19-37. 10.1016/0020-0271(71)90024-6 Retrieved from http://www.sciencedirect.com/science/article/pii/0020027171900246

da Costa Pereira, C., Dragoni, M., & Pasi, G. (2012). Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information Processing & Management, 48*(2), 340-357. 10.1016/j.ipm.2011.07.001 Retrieved from http://www.sciencedirect.com/science/article/pii/S0306457311000719

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics, 121*, 279.

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering, 10*(3-4), 327-348.

Gachet, A. (2004). *Building model driven decision support systems with dicodess* vdf Hochschulverlag AG.

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the Future, 2007*(2012), 1-16.

Gómez-Pérez, A., Martínez-Romero, M., Rodríguez-González, A., Vázquez, G., & Vázquez-Naya, J. M. (2013). Ontologies in medicinal chemistry: Current status and future challenges. *Current Topics in Medicinal Chemistry, 13*(5), 576-590.

Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., & Hofmann-Wellenhof, R. (2015). *Natural language processing and information systems* (2015th ed.). Cham: Springer International Publishing.10.1007/978-3-319-19581-0 Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-39146-0_2

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member Submission, 21*, 79.

Isaac, A., & Summers, E. (2009). SKOS simple knowledge organization system. *Primer, World Wide Web Consortium (W3C),*

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7*(5), 217-240.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv Preprint Cmp-Lg/9709008,*

Keen, P. G. (1978). *Decision support systems: An organizational perspective* Addison-Wesley Pub. Co.

Keen, P. G. (1980). Decision support systems: A research perspective. Paper presented at the *Decision Support Systems: Issues and Challenges: Proceedings of an International Task Force Meeting,* 23-44.

Khriyenko, O., Nguyen Kim, C., & Ahapainen, A. (2018). Cognitive computing supported medical decision support system for patient’s driving assessment. *GSTF Journal on Computing (JoC), 6*(1)

Klyne, G., & Carroll, J. J. (2006). Resource description framework (RDF): Concepts and abstract syntax.

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2006). Fedora: An architecture for complex objects and their relationships. *International Journal on Digital Libraries, 6*(2), 124-138.

Lally, A., Bagchi, S., Barborak, M. A., Buchanan, D. W., Chu-Carroll, J., Ferrucci, D. A., . . . Murdock, J. W. (2017). WatsonPaths: Scenario-based question answering and inference over unstructured information. *AI Magazine, 38*(2), 59.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database, 49*(2), 265-283.

Lin, D. (1998). An information-theoretic definition of similarity. Paper presented at the *Icml, , 98*(1998) 296-304.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association, 88*(3), 265.

Maguitman, A. G., Menczer, F., Roinestad, H., & Vespignani, A. (2005). Algorithmic detection of semantic similarity. Paper presented at the *Proceedings of the 14th International Conference on World Wide Web,* 107-116.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. Paper presented at the *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* 55-60.

Meystre, S., & Haug, P. J. (2006). Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics, 39*(6), 589-599.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM, 38*(11), 39-41.

NASA STI Program. (2012). NASA thesaurus. Retrieved from https://www.sti.nasa.gov/sti-tools/#thesaurus

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., . . . Chute, C. G. (2009). BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research, 37*(suppl_2), W173.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Power, D. J. (2002). Decision support systems: Concepts and resources for managers. *Studies in Informatics and Control, 11*(4), 349-350.

Power, D. J. (2014). Using 'Big data'for analytics and decision support. *Journal of Decision Systems, 23*(2), 222-228.

Prud, E., & Seaborne, A. (2006). SPARQL query language for RDF.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(1), 17-30.

Raimond, Y., Abdallah, S. A., Sandler, M. B., & Giasson, F. (2007). The music ontology. Paper presented at the *Ismir, , 2007* 8th.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv Preprint Cmp-Lg/9511007,*

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). Json-ld 1.0. *W3C Recommendation, 16*

Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web, 6*(3), 203-217.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. Paper presented at the *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics,* 133-138.

Appendix

A On Evaluating the Latest Prototype

The latest cloud-based version of the prototype can be accessed at:

https://oss.eu-gb.mybluemix.net/

The sources of the prototype are openly available at

https://git.eu-gb.bluemix.net/chinhnk/oss

Apart from the user interface, supported API requests are:

- **GET** */notes* - Get notes related to a patient name
    - **Query params**:
        - **patientName**: Name of the patient (eg: Matti)
    - **Response**: information related to patient
- **GET** */entities/:ontName/:id* - Get the ontology of an ID
    - **Path params**:
        - **ontName**: name of the ontology (eg: snomedct)
        - **Id**: the id of the ontology
    - **Response**: jsonLD response of the ontology
- **POST** */configs/:configType* - Change the config of the base doc or target doc
    - **Path params**: configType: configtype: **base** or **target**
        - **Request body**: fieldname: configFile, json config file of the target file.

  - **Response**: if successful, response back the response file

- **POST** */targetDoc/upload* - Upload target document
  - **Request body**:
    - **field:** collectionName: string name of the collections
    - **field**: targetDocs (file): target document, can be one or multiple files
  - **Response**: if successful, response back the response file
- **POST** */baseDoc/upload* - Upload base document
  - **Request body**:
    - **field**: docName: name of the base document
    - **field**: baseDocs (file): the base document file in .txt format

Response: if successful, Response status 200