Cognitive Computing supported Medical Decision Support System for Patient's Driving Assessment

O. Khriyenko, C. Nguyen Kim and A. Ahapainen

Faculty of Information Technology, University of Jyväskylä, P.O. Box 35, FIN-40014 Jyväskylä, Finland
oleksiy.khriyenko@jyu.fi, chinhnk.93@gmail.com and atte.ahapainen@hotmail.com

Abstract— **To smartly utilize a huge and constantly growing volume of data, improve productivity and increase competitiveness in various fields of life; human requires decision making support systems that efficiently process and analyze the data, and, as a result, significantly speed up the process. Similarly to all other areas of human life, healthcare domain also is lacking Artificial Intelligence (AI) based solution. A number of supervised and unsupervised Machine Learning and Data Mining techniques exist to help us to deal with structured data. However, in a real life, we pretty much deal with unstructured data that hides useful knowledge and valuable information inside human-readable plain texts, images, audio and video. Therefore, such IT giants as IBM, Google, Microsoft, Intel, Facebook, etc., as well as variety of SMEs are actively elaborating different Cognitive Computing services and tools to get a value from unstructured data. Thus, the paper presents feasibility study of IBM Watson cognitive computing services and tools to address the issue of automated health records processing to support doctor's decision for patient's driving assessment.**

Keywords - ***cognitive computing; decision support system; medical record processing; natural language processing; semantic similarity; IBM Watson; medical ontology; driving assessment.***

I. INTRODUCTION

Nowadays, we are dealing with a huge and constantly growing volume of data. To smartly utilize all the collected data, improve productivity and increase competitiveness in various fields of life, human requires decision making support systems that efficiently process and analyze the data and significantly speed up this process.

Taking into account rapid digitalization of societies and businesses, decision making support systems are based on dynamic service oriented and semantically facilitated infrastructures that seamlessly integrate heterogeneous data from various sources. Further, to support a decision making based on integrated and aggregated data, system has to apply variety data processing, optimization, automated knowledge retrieving and inferencing techniques [1]. A knowledge-driven decision support system (KD-DSS) provides specialized problem solving expertise stored in some knowledge representation format like facts, rules, procedures, or in similar structures and it suggests or recommends actions to human expert. Advanced analytical tools like data mining are integrated within the KD-DSS to find hidden patterns in information retrieval and knowledge discovery processes [2]. A number of supervised and unsupervised machine learning and data mining techniques exist to help us to deal with structured data. But in a real life, we pretty much deal with unstructured data as well. It is a data, where powerful knowledge and valuable information are hidden inside human-readable plain texts, images, audio and video materials [3][4][5][6]. To address this type of data, machines have to be able to understand it not only on syntactical, but, what is even more important, on a level of semantics that data contains. Therefore, big players such as IBM, Google, Microsoft, Intel, Facebook, etc., as well as variety of SMEs are working on elaboration and offer different Cognitive Computing[1] services and tools that heavily utilize Natural Language Processing[2] (NLP), Semantic Web[3] and Linked Data[4], and Deep Learning[5] technologies to get a value from unstructured data to be useful in decision making support.

With respect to application areas and domains where decision making support systems would be useful and helpful, the scope is unlimited. It could be any area, where we have a huge amount of logged data and profiles, collected feedbacks and observation measurements, data produced by IoT devises and social web content produced by human, human driven statements and guidance, mass media content or other data that could be processed to retrieve and infer hidden knowledge. Businesses and industries, health and social care, education and science, sport and wellbeing, security, transportation and other areas of human life are lacking AI based solution to facilitate decision support.

This paper tackles the problem of medical decision support system in general and assessment of driving capabilities of the patient in particular. Regarding to the use-case, brought to the project by medical experts, there is a need for supportive system that automates the process of relevant information filtering to speed up a patient assessment. With respect to the case scenario the main duty of a doctor is to recognize a person

[1] https://en.wikipedia.org/wiki/Cognitive_computing
[2] https://en.wikipedia.org/wiki/NLP
[3] https://en.wikipedia.org/wiki/Semantic_Web
[4] https://en.wikipedia.org/wiki/Linked_data
[5] https://en.wikipedia.org/wiki/Deep_learning

DOI: 10.5176/2251-3043_6.1.114

who is not eligible to have a driving license due to his/her health condition. Decision is made based on a health profile of a person that consists of health statements/records regarding various diseases that person has. To make a problem definition is more narrowed and concrete, we introduced several assumptions on top of the original case. Currently, patient profile is distributed among different systems making it more difficult for doctor to process the data. Since, there are ongoing activities at the hospital aimed at integration of the systems, the first assumption we made is that we are dealing with a simple integrated access point to retrieve health profile of a patient. Second assumption is that each health statement consists of two parts: structured information (e.g. disease, various lab test results and measurements, prescribed medications, etc.), and unstructured part in a form of plain text that contains doctor's comments with valuable for decision making information. With respect to the expert's point of view, one single diagnose doesn´t necessarily affect driving capability and doctors decision; but information (symptoms, signs, patient behavior, etc.) described in doctor´s comments gives valuable key elements for actual decision.

Therefore, the main objective of this work is to make a feasibility study of existing IBM Watson cognitive services and check their applicability to tackle the mentioned problem. Paper presents several approaches for decision support system created based on the tools facilitated by IBM Watson technology, as well as comparison of their performance. These approaches are described in the Section 2. Section 3 presents prototype development details and comparison results. Further sections refer to related works, conclude current work and guide towards future achievements.

II. KNOWLEDGE TRANSFER

When we are talking about artificial intelligence and decision support systems in particular, we are trying to delegate some duties/functionality from human to machine/software. As a result, we refer to the way we transfer the knowledge into certain model that will be used by the system. With respect to our use-case, the knowledge required for decision making is available from two sources:

- "mind" - expert knowledge of a doctor
- "book" - instructions written in a human readable form (e.g. plain text)

The main objective of the use-case is to elaborate a decision support system that is able to automate a process of identification people with health-based restrictions to have a driving license. Here we may consider several levels of system performance: from simple classification of the statements with certain probability to contain data valuable for decision making, towards more advanced performance with justification of made decision. All of them require relevant knowledge being presented in some machine-readable form depending on a chosen approach. Thus, we have to consider options of knowledge transfer from initial sources ("mind" or "book") into "machine" - machine readable representation (e.g. rules, neural network model, set of relevant entities and keywords, etc.). Such transformation could be done manually or be automated (or semi-automated).

Manual "mind-to-machine" or "book-to-machine" transfer assumes that domain expert generates corresponding decision making rules, populates a set of relevant entities, etc. based on his/her knowledge or knowledge hidden in the human readable instructions/guides. Manual approach is time consuming and causes huge workload for expert. Automated transfer requires advanced data processing logic and/or adopted machine observation environment to retrieve the knowledge from the natural language or from observation of expert's behavior. In this case, semi-automated approach is meant for the cases when corresponding algorithm cannot automatically make transformation with high enough confidence and still requires final tuning or confirmation from human.

Within our prototype we aimed at elaboration of decision support system with automated/semi-automated functionalities facilitated by IBM Watson[6] cognitive computing services and other tools. Implementation does not involve any actual decision making, rather provides a supportive functionality that processes a huge amount of samples (patient statements) and filter/select the most relevant ones with respect to a set goal. Depending on a knowledge source we have, we present following two approaches that could be used separately, as well as be combined.

A. *Book-to-Machine knowledge transfer based approach*

To estimate a level of relevance of certain health record/statement to the case of ineligibility to have a driving license, system should recognize accordance of the patient health statement/record to the subject knowledge. Our original idea was to use Watson Natural Language Understanding[7] (NLU) cognitive service to extract hidden rules in regulatory documents. In our case, such a control document is the translated version of "Ajoterveyden arviointiohjeet lääkäreille" - "Guidelines for doctors assessing fitness to drive" issued by Trafi[8] (Finnish Transport Safety Agency).

Watson NLU provides the capability to analyze unstructured text documents for categories, concepts, keywords, semantic roles, entities, relations, as well as emotion and sentiments. The service includes a default general domain language model which can categorize documents into 1 083 categories and recognize up to 24 entity types, 433 entity subtypes and 53 relation types. Watson NLU service makes possible to retrieve a semantic triple(s) (subject-predicate-object) based on "Relations" and "Semantic Roles" generated from analyzed text input. "Relations" recognizes when two entities are related, and identify the type of relation. "Semantic Roles" parses sentences into subject, predicate, and object form. Extracting semantic triples from the control document, we populate RDF[9] storage and build a control knowledge space that will be queried against triples extracted from patient statements (their structured and unstructured parts). However, based on experimental results we may conclude that default general model is not fine-tuned for medical domain, the

[6] https://www.ibm.com/watson/

[7] https://www.ibm.com/watson/services/natural-language-understanding/

[8] https://www.trafi.fi/en/about_trafi

[9] https://www.w3.org/RDF/

semantic roles analyses are not accurate enough for reliable rule extractions yet. It is still possible to improve the extraction performance by connecting customized medical domain oriented model in conjunction with the default model for domain-specific analyses.

In order to create a custom model for Watson NLU, IBM offers Watson Knowledge Studio [10] (WKS) - a stand-alone product that aims to better involve field-experts in the training of supervised machine learning models in order to process unstructured data. The product offers a user-friendly interface and features which enable collaboration through iterative processes. Nevertheless, creation of sophisticated healthcare domain language model with WKS requires comprehensive analysis of problem domain from knowledge management expert, as well as time consuming affords from medical experts to improve the model with extra supervised machine learning based facilitation.

Therefore, we resorted to building a custom model which, in the current initial version, focuses on medical entities (diseases, symptoms, medications, lab test results and other measured indicators, and any other relevant to the subject "keywords") recognition. Now we have to define a space of control entities - entities the most relevant to our case/subject. To create such a set of control entities, we use our control document to extract all the "Entities" and "Keywords" using Watson NLU service and keep them as a control set of entities. To improve performance and get better result comparing to default language model, we facilitated outcome of the service by using custom domain specific model in addition. We may populate the model with domain specific entities by using existing external sources with a set of healthcare related entities (such as official disease classification, medication names, etc.). Among such sources we may highlight ICD-10 [11] and SNOMED CT [12]. ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. SNOMED CT is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. It is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. So far, WKS provides a graphical user interface for knowledge manager to create the model and does not support a programmable way to do that. Fortunately, WKS allows uploading existing sets of entities or other kind of vocabularies into the model if the sets are presented in one of supported formats (e.g. CSV, JSON, etc.). This way we may still use existing external sources to create a domain specific custom model. Additionally, we may create and add any set of relevant abbreviations which are heavily used by doctors in their notes/statements. Here we may use pattern (regular expression) based approach to automatically retrieve abbreviations for a text, but, to have meaningful explanation of them; we require manual involvement of domain experts.

The evaluation assumption for relevance of a health statement/record to the case of driving assessment (ineligibility to have a driving license) is following - as more entities from patient record happen to be common with entities from domain specific control document, then more relevant the record to the case (see Fig.1). There, $N_i^{e'}$ and N_i^e are amount of entities common for patient record PR_i and control document cD, and overall amount of entities detected in PR_i correspondently. Setting corresponding threshold for the density, we may eliminate irrelevant records from further analysis.

However, taking into account that doctor's notes in patient's health records might contain several subjects/topics written in different paragraphs, we do analysis on paragraph basis and consider the record relevant if any of the paragraphs (sub-records) pass the density threshold.

To be more precise, we may take into account not only exact match of the entities or their synonyms (that are considered as equivalent entities with similarity coefficient 1), but also consider semantic similarity of them. Therefore, each entity has to be accompanied with semantic similarity weight (which might be cut by certain threshold to eliminate influence of entities with low relevance). Thus, an amount of entities common for patient record PR_i and control document cD is calculated as a sum of semantic similarity weights (above a threshold) of entities from PR_i with respect to the entities of cD ($N^{e'} = \sum w^s$). If there are several semantic similarity weights (above threshold) for the same entity from PR_i, the entity is counted only once with bigger weight. Semantic similarity of the entities is calculated by Google Similarity Distance [7] applied on top of domain knowledge aggregated under medical domain ontologies: SNOMED CT, FMA[13] and NCI[14]. The set of text samples/documents was generated from textual descriptions of the concepts present in the ontologies under description properties *rdsf:label* and *rdfs:comment*.

As soon as we detect relevance of the patient record to the subject, we have to support doctor with the reference to the most relevant part of the control document associated with the record (or the record's part). Based on density of common

$$\rho_{PR_i} = \frac{N_i^{e'}}{N_i^e} \quad ,i = \overline{1,n}$$

Figure 1. Relevance to the subject through density of subject related entities.

[10] https://www.ibm.com/us-en/marketplace/supervised-machine-learning

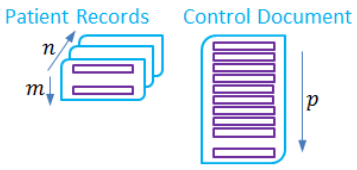
[11] https://en.wikipedia.org/wiki/ICD-10, http://www.who.int/classifications/icd/en/

[12] https://en.wikipedia.org/wiki/SNOMED_CT, http://www.snomed.org/snomed-ct/what-is-snomed-ct

[13] http://si.washington.edu/projects/fma

[14] http://www.obofoundry.org/ontology/ncit.html, https://ncit.nci.nih.gov/ncitbrowser/

$$R_{ijk} = \sum_{l=1}^{|E'_{ij}|} \left| \rho_l^{PR_{ij}} - \rho_l^{cD_k} \right| \quad , i = \overline{1,n} \quad j = \overline{1,m} \quad k = \overline{1,p}$$



$$, where \ \rho_l^{PR_{ij}} = \frac{N_{e'_l}^{PR_{ij}}}{N_{e'}^{PR_{ij}}} \ and \ \rho_l^{cD_k} = \frac{N_{e'_l}^{cD_k}}{N_e^{cD_k}} - are \ densities \ of common \ entity \ in \ PR_{ij} \ and \ cD_k$$

$$correspondently, \ E'_{ij} = \{e'_1, e'_2, \dots, e'_l, \dots, e'_L\} - set \ of \ entities \ common \ for \ PR_{ij} \ and \ cD \ .$$

Figure 2. Density based relevance distance.

entities we may calculate a similarity distance (see Fig.2) of the records to the different parts of the control document and order them to facilitate further doctor's decision making. Density $\rho_l^{PR_{ij}}$ of a common entity e'_l in patient record PR_{ij} is a ratio of number of appearance of the entity in the record and number of all common entities (including repeated entities) in the record. Similarly, density $\rho_l^{cD_k}$ of the same entity e'_l in a part of control document (sub-document) cD_k is a ration of number of appearance of the entity in the sub-document and number of all entities (including repeated entities) in the sub-document.

Alternatively to the present approach, we may also use "retrieve" functionality of IBM Watson Retrieve and Rank service (R&R) that was transformed into Discovery[15] service to find relevant to the case health statement(s). Having a possibility to get a feedback from a system user (doctor) in a form of relevance ranking, "rank" functionality of the service could be also used to improve the service results. Since, Watson R&R service works with textual samples, structured part of a statement/record should be transformed into a text and integrated with unstructured part. It could be done via simple transformation pattern (e.g. "patient has <disease name>", or "sugar level is <amount and units>", etc.). Using Watson Discovery service, we present our control document(s) as a cluster of sub-documents among which we try to find one relevant to a query, which is a text based representation of a health record. Thus, if a record contains information which is relevant to any part from a control document, the statement is considered as relevant to our case and will be suggested for doctor to be checked.

B. Mind-to-Machine knowledge transfer based approach

Alternatively to manual formalization of human knowledge and experience, automated approach assumes existence of an environment through which system is able to observe human's (doctor's) actions and learn from it. This way we transfer knowledge that doctor has into machine in more or less unobtrusive for user way. In our use-case, such an environment could provide a graphical user interface through which doctor can browse patient's health statement/record and able to highlight the parts that are considered as valuable/important for the decision (s)he makes. Thus, dataset, collected by observation environment, consists of highlighter by doctor *key:value* pairs from structured part of a statement and highlighted text from unstructured part of it, as well as decision (eligible or not) made by doctor as a label for the set. Having labeled training set as a result of decision making process observation, we may train classification model(s) to be further used for decision support in processing of new statements/records.

For classification of the structured part of the statement, we may consider several Machine Learning models (such as Decision Tree, Neural Networks, etc.). Since use-case prototype development was focused on added value that can be brought by processing of unstructured part of the patient health statements, we integrated structured part into unstructured as a text (following the same pattern based transformation approach described above) and used IBM Watson Language Classifier[16] (LC) services as an integrated solution. Watson Language Classifier allows us to train classification model and classify any new input into one of two classes: relevant (ineligible) and irrelevant (eligible) to our case. As training examples for "relevant" class we used text-based parts of the statements collected during observation. For the second class, training samples come from statements that were not considered by doctor as cases where patient ineligible to have a driving license.

III. PROTOTYPE IMPLEMENTATION

This section presents the architecture and implementation of the prototype with respect to two "Book-to-Machine" knowledge transfer approaches described above. Since we have faced legacy and privacy issues regarding patient personal data, we did not manage to collect big enough training set of anonymized patient records to apply classification approach based on IBM Watson Language Classifier service. Following, we present overviews and architectures of the developed solutions, limitations of the existing Watson-technologies that were found during the development, comparative analysis of the solutions and proposals for further improvement.

A. Solution based on medical entity enriched custom language model

As mentioned before, the solution is a general decision support system for the doctors to process the unstructured portion of the doctor's notes more efficiently. It highlights

[15] https://www.ibm.com/watson/services/discovery/

[16] https://www.ibm.com/watson/services/natural-language-classifier/

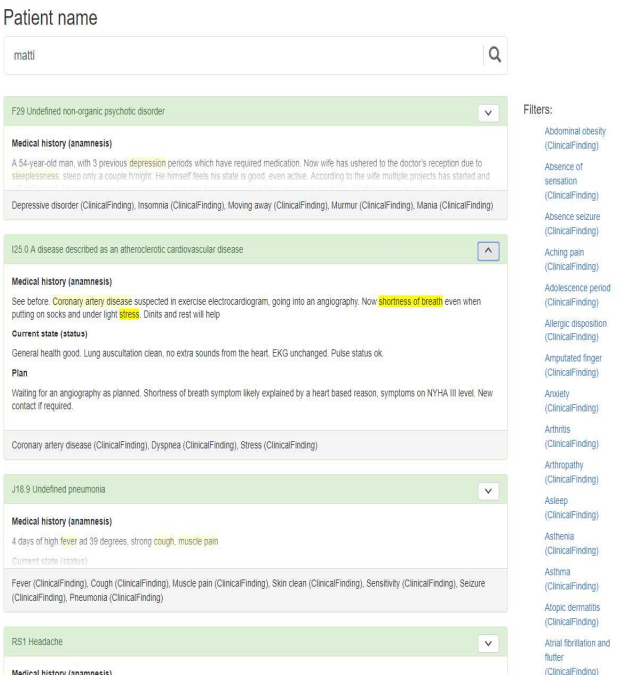


Figure 3. Doctor's notes view.

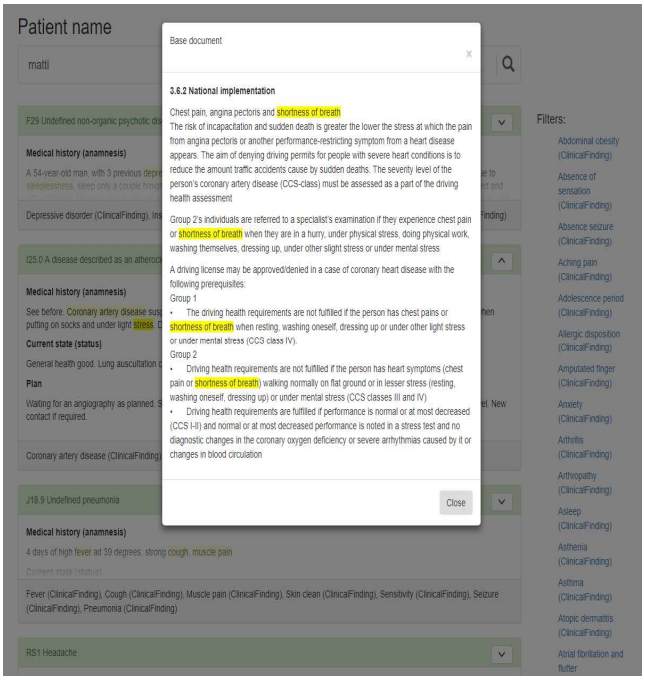


Figure 4: Control document excerpts.

keywords from the notes and shows the relation between them and the control document. The prototype has a search function to search patient records and doctor notes using a patient's name (see Fig.3). After a search is made, the search results are shown under the search input area as result cards. These cards consist of a header which is the main diagnosis, a body that contains patient's medical history, the current state of the patient and a plan to proceed with. At the bottom of the card there are tags that consist of the entities and their types that are found in the result card via Watson NLU service. Next to the search results there is a list of filters that can be used to filter the result cards by their tags. The filters only include subject domain related entities which can be found in the base document.

In the real-life case, system will go through all unanalyzed patients' records and sort them based on estimated relevance level of the record to our Driving Assessment case. Any newly created record will be also analyzed and placed in correspondent order with other records. Thus, system will prepare the most promising (most likely relevant) records for doctor's assessment. Parts of the text are highlighted if they are recognized as an entity by the solution. Text highlighted with bright yellow indicates entities that have related control document excerpts. Upon clicking on the highlight, a modal dialog box containing related control document excerpts opens. The control document excerpts also have highlighted parts that are related to the entity that was clicked (see Fig.4).

As the development environment for the prototype we used *IBM Bluemix* cloud. As a *platform as a service (PaaS)*, IBM Bluemix already handles the tedious tasks of setting up and configuring infrastructures and provides developers with ready-made *boilerplates* and *continuous delivery toolchains*, which greatly speed up the development and deployment processes. It also emphasizes service-oriented architecture with a wide range of functionalities such as *storages*, *analytics*, *IoT*, etc. - all available as web services. Our prototype uses *Node.js Cloudant DB Web Starter* boilerplate from IBM Bluemix[17] and two other Bluemix services, which are:

- IBM Watson NLU;
- Cloudant NoSQL DB.

At the center of the solution (see Fig.5) is the IBM Watson NLU cognitive service. Each HTTP request to the service API contains a chunk of text to be analyzed or a URL to a webpage, the textual content of which will 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 use Watson NLU via its *Node.js* SDK.

In order to create a *custom model* for Watson NLU, we used WKS. 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 our prototype, we have focused on constructing an extensive collection of dictionaries and deploy it via a rule-based annotator. These dictionaries can later be used to pre-annotate

[17] https://www.ibm.com/cloud-computing/bluemix/

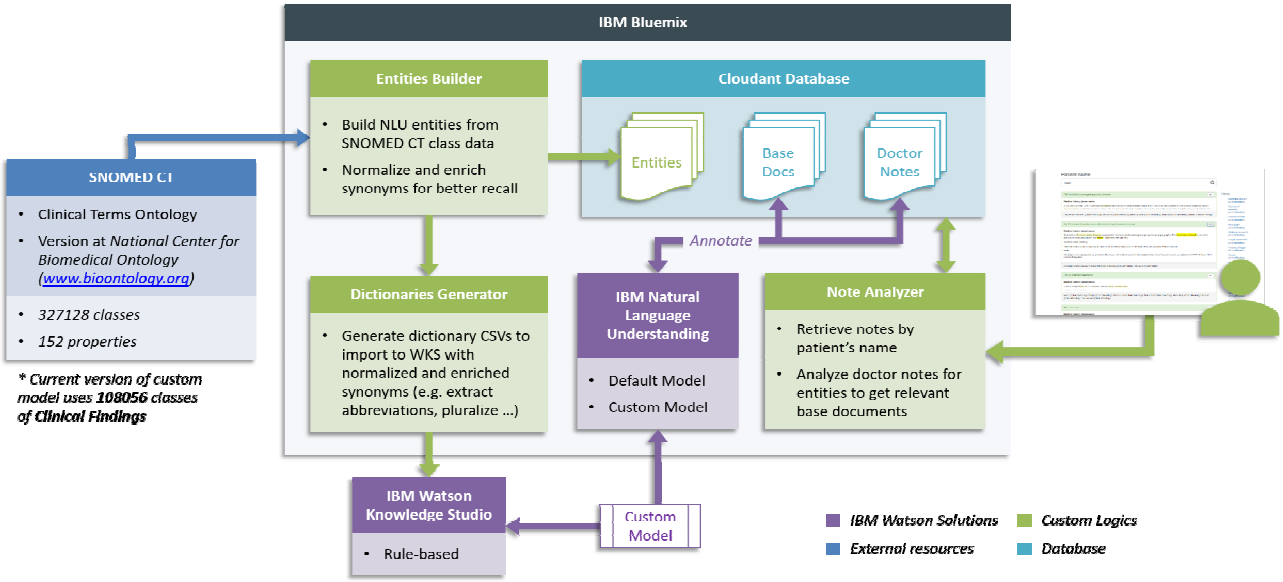


Figure 5. Solution Architecture.

documents in order to assist human annotators with a set of preliminary annotations. 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 in order 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.

Our dictionary entries are generated based on class data of SNOMED-CT ontology available on BioPortal[18]. SNOMED-CT is adopted as a standard for electronic health information exchange by the United States and the United Kingdom[19] and is used by more than fifty countries worldwide. BioPortal is an integration of ontology services and related tools by The National Center for Biomedical Ontology, funded by the US National Institutes of Health. Its features include a REST API which can be used to query its collection of 593 aligned ontologies. Following Semantic Web and Linked Data standards [8][9], the SNOMED-CT ontology is an RDF representation of SNOMED-CT, containing 327128 classes and 152 property types. Since the main focus of our prototype is to detect medical entities, we first made use of data of 108056 subclasses of *Clinical Finding (http://purl.bioontology.org/ontology/SNOMEDCT/404684003)*.

The Entities Builder module in our prototype was made to achieve two objectives: (1) prepare and enrich data for Dictionaries Generator; and (2) 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 text that has higher case, therefore all of the surface forms except abbreviations should be in lower-case.

Entity analyses done with Watson NLU default model can provide disambiguation when applicable, where 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. In our current version of the prototype, all entities share a common main type - *ClinicalFinding* - and their subtypes are the names of the direct subclasses of the *Clinical Finding* class of which they are descendants.

B. *Discovery (Retrieve and Rank) based solution*

This section presents tests that were conducted in relation to driving assessment case using IBM Watson Discovery cognitive service. Watson Discovery search service is built over Apache Solr[20] search engine. Solr is a powerful open source enterprise search engine which is designed for querying huge amounts of data in fast and efficient manner. Discovery service uses the Solr for the initial query and then the service ranks the query results with respect to the trained ranking model. To conduct queries and rank the results, the service needs documents to query from. For this purpose, service uses

[18] http://bioportal.bioontology.org/ontologies/SNOMEDCT
[19] https://www.nlm.nih.gov/medical-terms.html
[20] http://lucene.apache.org/solr/

$$D = \frac{\sum_{i=1}^{n} k * \left|R_i^S - R_i^H\right|}{D_{max}}$$

$$, where\ k = \begin{cases} 1, & if\ result\ unit\ belongs\ to\ the\ set\ of\ result\ units\ of\ the\ solution \\ 0, & otherwise \end{cases}$$

$$n - number\ of\ different\ result\ units\ in\ both\ solutions,$$

$$R_i^S and\ R_i^H - ranks\ of\ the\ result\ unit\ in\ solution\ and\ human\ ranking\ correspondently,$$

$$D_{max} \in \{12,15,18,21,24,28\} - maximum\ distance\ with\ respect\ to\ number\ of\ result\ units\ (n).$$

Figure 6. Normalized distance between decisions made by solution and human.

IBM Watson Document Conversion [21] service that divides given Word, PDF or HTML documents into answer units. An answer unit can be a single document or a portion of document. Document Conversion service uses headers in the document to divide the document into answer units if the user chooses to do so. The documents are added into a collection of documents that is used for queries. To be used, the Discovery service needs to be initially trained. This initial training will create "ground truth" that the ranker will use when ranking the query results. To create the ground truth the service requires 150 sample queries and for the user(s) or the domain expert(s) to rank the query results for each query by hand. This is done using a star rating system (**** = spot on answer, *** = partially answers the query, ** = doesn't answer the query but is relevant to the topic, * = incorrect answer). After the ground truth is created the service is ready to process queries. The service can also be improved with additional training and it also automatically generates tasks that aim to improve the service performance for the user. All of this can be easily done using the graphical tooling interface provided by the service, although more complex actions require some coding skills (for example tweaking the settings etc.).

Due to insufficient (limited) amount of data samples (anonymized patient records), training of service "ranker" did not show actual improvement in comparison to the results of service "retriever". Therefore, we decided to test "retrieve" functionality of the service based on the same data samples (patient records) and the control document as in the previous solution (Section 3a). Similarly, we split the control document (Trafi document) to the Watson Discovery service's document collection - set of document's paragraphs that are used as answer units. Each result of the service query consists of five different answer units. The answer units come in order of relevance to the query. Unfortunately, they come without any relevance distance value or confidence level that could be used for more precise comparison.

C. *Solution comparison*

In this section we compare performance of our solution that is based on medical entity enriched custom language model and performance of Watson Discovery service based approach. To equally test the solutions and normalize the results, we do the comparison based on first five results, since Watson Discovery based solution returns only that amount. It would be much precise comparison if we use actual relevance measure or confidence level of the results. However, Watson Discovery based solution provides only simple ranks (1st,2nd,3rd,4th,5th) for them. Thus, we have to disregard more precise distribution of relevance distances of the first solution, and take into account only the rank/order of the result.

Assuming, that first five results from two solutions might not consist of the same result units, the size of result set for each query may vary from 5 up to 10 result units. Since we would like to check which of the solutions produces result similarly to what human (expert) does, the result set of both solutions is evaluated by human and their result units are ranked in relevance order for human point of view. Further, distance between the granted ranked/places is converted to the performance measure for each of the solutions (see Fig.6).

Due to the privacy related issues we had a limited amount of anonymized patients' records for test purposes, and processed only 40 patient records. Average closeness of automated solution that is based on medical entity enriched custom language model to the decision made by human was 92%. In contrast, Watson Discovery service based solution did show quite poor result in 67%. But, here we had limitations and did not use "rank" part of service functionality.

IV. Related Work

Regarding related work, we would like to mention WatsonPaths [22] and Watson EMR Assistant - new cognitive computing projects that enable natural interaction between physicians, data and electronic medical records. Technologies are expected to help physicians make more informed and accurate decisions faster and to cull new insights from electronic medical records (EMR). The projects will create technologies that can be used by Watson in the domain of medicine. Unfortunately, currently we do not have a possibility to try their solution out yet, and cannot make actual comparison with the solution we present.

With respect to transformation of unstructured data into structured machine-processable form of RDF triples, we would highlight user-driven semi-automated approach presented in [10][11], where author tackles the problem of free text based customer feedback handling for further automated processing. In our opinion, use of Watson NLU service for automated semantic triples extraction would be suitable approach in case we have a comprehensive domain oriented custom language mode.

[21] https://www.ibm.com/watson/services/document-conversion/

[22] http://www.research.ibm.com/cognitive-computing/watson/watsonpaths.shtml#fbid=VGlW8Qpyjte

V. Conclusions

This paper presents the approach we used to elaborate decision support system for medical experts in the context of driving assessment based on patient records. In the context to the mentioned limitations and restriction with respect to the defined problem, we consider achieved results reasonably good (that was also has been acknowledged by medical expert participated in the project). In our approach we apply density based similarity distance calculation with a purpose to not only take into account amount of common entities, but also put more value to similarity of relevant entities density. Similarly, instead of simple Jaccard distance[23] it might be reasonable to apply Angular distance[24].

Named-entity recognition is only the first step in achieving our original goal. Since we conclude that the named-entity recognition feature of the prototype has been adequate, we are ready to move on to relation recognition which involves detecting patterns of relationships between entities and map these relationships with Semantic Web compliant relations (triples). To facilitate *custom language model*, this task will involve establishing a more complete type system for entity relationships in WKS, exploring existing methods of relation mapping in natural language processing as well as deriving original solutions if would be needed. We also see reasonable to analyze conditional patterns for rule extraction. We would like to investigate the works of [12][13][14] and try to implement the presented pattern matchings using WKS. With respect to the medical domain, it would be reasonable to make a feasibility study on applying Unified Medical Language System[25] (UMLS) as integration solution for huge variety of vocabularies in medical domain. In current solution we have not made extensive use of WKS machine learning annotator due to limited medical expert human resource. Therefore, since our initial result has positively proven the concept, we would like to further refine our solution and will attempt to train a WKS machine learning annotator. It is also worth mentioning that IBM Watson solutions have been constantly evolving, which means periodical reassessments of their capabilities are advised. Finally, gathering more sufficient amount of labeled data to be used as a training set for classification model, we will test Watson Language Classifier service as was mentioned in the Section 2b.

Even though presented approaches were considered in the context of driving assessment and have been bound to the Trafi regulation document (control document) as a "book" source of knowledge, they might be applied to any similar use-case were domain knowledge is collected in human readable natural language within a control document.

[23] https://en.wikipedia.org/wiki/Jaccard_index

[24] https://en.wikipedia.org/wiki/Angular_distance

[25] Unified Medical Language System: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

Acknowledgment

The research is done in the "VALUE FROM HEALTH DATA WITH COGNITIVE COMPUTING" project led by Faculty of Information Technology (University of Jyväskylä) in collaboration with IBM, Central Finland Central Hospital and other medical/social care organizations. Authors are especially thankful to doctor Heikki Alanen and other project team members for fruitful collaboration and useful discussions.

References

[1] M. H. Murtadha and A.Q. Banaz, "Knowledge-Driven Decision Support System Based on Knowledge Warehouse and Data Mining for Market Management". International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 1, January 2014, ISSN 2319 - 4847

[2] A. Souza Inácio, R. Andrade, A. Wangenheim and D.D.J. Macedo, "Designing an information retrieval system for the STT/SC", *e-Health Networking Applications and Services (Healthcom) 2014 IEEE 16th International Conference on*, pp. 500-505, 2014.

[3] A.R. Al-Azmi, "Data, Text, and Web Mining for Business Intelligence: a survey", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, March 2013.

[4] J. Friedlin, M. Mahoui, J. Jones and P. Jamieson, "Knowledge Discovery and Data Mining of Free Text Radiology Reports". In Proceedings of the First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB), San Jose, CA.

[5] S. Keretna, C.P. Lim and D. Creighton, "A hybrid model for named entity recognition using unstructured medical text". In Proceedings of the 9th International Conference on System of Systems Engineering (SOSE), Adelaide, Australia, June 9-13, 2014 pp. 85-90.

[6] M. Bogatyrev and K. Samodurov, "Knowledge Discovery from Texts with Conceptual Graphs and FCA". In Proceedings of International Workshop on Formal Concept Analysis for Knowledge Discovery (FCA4KD 2017), Moscow, Russia, June 1, 2017.

[7] R.L. Cilibrasi and P.M.B. Vitanyi, The Google Similarity Distance. In: IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No 3, March 2007, 370–383

[8] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", Scientific American 284(5),pp.34-43.

[9] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space" (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.

[10] O. Khriyenko, "Customer Feedback System: Evolution towards semantically-enhanced systems", In: Proceedings of the 11th International Conference on Web Information Systems and Technologies, WEBIST 2015, 20-22 May, 2015, Lisbon, Portugal, 518-525.

[11] O. Khriyenko, "Customer Perception Driven Product Evolution: Facilitation of Structured Feedback Collection.", In: Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST 2016), 23-25 April, 2016, Rome, Italy, pp. 196-203.

[12] T.D. Breaux and A.I. Anton, "Mining rule semantics to understand legislative compliance". In: Proceedings of the 2005 ACM workshop on Privacy in the electronic society (WPES '05). Alexandria, VA, USA — November 07 - 07, 2005, pp. 51-54

[13] T.D. Breaux and A.I. Anton, "Analyzing regulatory rules for privacy and security requirements". IEEE Transactions on Software Engineering, 34(1):5–20.

[14] T.D. Breaux, M.W. Vail and A.I. Anton, "Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations." In 14th IEEE International Requirements Engineering Conference (RE'06), pages 49–58. Institute of Electrical and Electronics Engineers (IEEE), September, 2006.

Oleksiy Khriyenko (1981) obtained Engineer's degree in Computer Science (Intelligent Decision Support Systems) in 2003 from the Kharkov National University of Radioelectronics, Ukraine. Later, Oleksiy Khriyenko obtained a Master's degree in Mobile Computing from MIT department (University of Jyväskylä, Finland). Since 2008 he is Ph.D. from the same department. His research interests include: Artificial Intelligence, Deep Learning and Cognitive Computing, Semantic Web and knowledge engineering, multi-agent systems, Web of Things and ubiquitous services, context-sensitive adaptive environments, etc. Currently, Oleksiy Khriyenko does research, lecturing and is involved in management of international master programs (WISE, COIN) at IT faculty, University of Jyväskylä. (http://users.jyu.fi/~olkhriye)

Chinh Nguyen Kim (1993) graduated with a Bachelor of Computer Science Degree in Vietnam in 2015 with the focus of Machine Learning and Natural Language Processing (NLP). He is currently pursuing a Master's Degree of Web Intelligence and Service Engineering in University of Jyväskylä, Finland. In 2017, he took part in the "Value from Public Health Data with Cognitive Computing" and "Health Cloud" projects at the University as a research assistant, in which he focused on evaluating the application of Semantic Technologies and NLP in processing and generating values from healthcare data.

Atte Ahapainen is a master's degree student in the Faculty of Information Technology at the University of Jyväskylä. In 2017, he worked as a research assistant in the "Value from public health data with cognitive computing" and "Health cloud" initiatives at the University and is currently a software developer in the incubator team at Solteq plc.

