Note Highlights: Surfacing Relevant Concepts from Unstructured Notes for Health Professionals

Vanessa Lopez, Joao Betterncourt-Silva, Grace McCarthy,
Spyros Kotoulas, Natasha Mulligan, Fabrizio Cucci,
Stephane Deparis, Marco Luca Sbodio, Pierpaolo Tommasi
IBM Research, Dublin, Ireland

John Segrave-Daly, Conor Cullen, Ciaran Hennessy,
Beth McKeon, Nagesh Yadav, Karie Kelly,Russell Olsen
IBM Watson Health, Dublin, Ireland

John Dinsmore, Anne-Marie Brady,
School of Nursing & Midwifery, Trinity College Dublin, Ireland

***Abstract*— Health and social care professionals are under increasing pressure to assimilate the ever-growing volume of data from case notes and electronic medical records. In this paper, we propose and evaluate with domain experts a cognitive system for patient-centric care that leverages and combines natural language processing, semantics, and learning from users over time to support care professionals making informed and timely decisions while reducing the burden of interacting with large volumes of unstructured patient notes. We propose methods for highlighting the entities embedded in the unstructured data and providing a personalized view of an individual. We evaluate through a user study and show a consensus between what the domain experts and the system consider relevant and discuss early feedback on the value of our Note Highlights methods to domain experts.**

Keywords—Natural Language Processing; Health Information Systems; Artificial Intelligence

I. INTRODUCTION

Complex cases in health and social care account for the majority of healthcare costs worldwide. The key drivers behind these growing costs are the burden of chronic diseases and the increase in prevalence of multimorbidity [2]. In light of this, the appropriate management of complex cases has become one of the most important challenges for health systems [2]. Current approaches that focus on a single disease should be complemented by the work of generalists, providing continuity, coordination, and individualized care for multimorbidity patients [2]. This has led to a new paradigm where "mutually dependent" [9] multidisciplinary teams, physicians, nurses, social care workers, and informal caregivers are, among others, combining their expertise to gain a holistic view of the patient and to deliver tailored care that enables better outcomes [19][20].

The adoption of electronic health records and other technologies that support the delivery of care have contributed to a growing volume of data while promising to improve quality of care and reduce costs [25]. However, healthcare professionals now have to cope with the burden of trawling through large numbers of case notes that are poorly structured, not easily accessible, and that do not provide adequate support for decision-making [26]. Essentially, there is a shift from a situation where not enough information is shared to one where the vast volume of information shared becomes a burden in itself. Research in health informatics has focused on tackling some of these challenges - yet available systems are still not able to fully satisfy user needs [16]. New methods and systems are needed to highlight the most pertinent information and provide relevant insights to practitioners at the point of care [22].

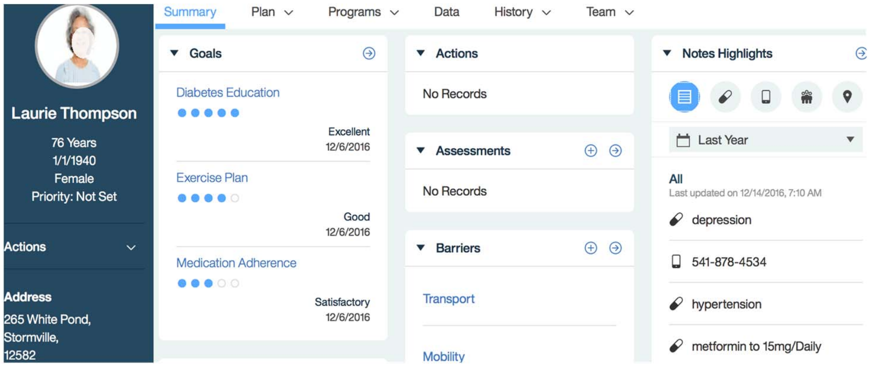
This paper presents a system that supports care teams in collecting the right multi-disciplinary information across multiple sources, helping them making informed and personalized decisions. This is achieved by adopting a cognitive computing approach [6] that leverages natural language processing tools over unstructured text, semantic technologies for incremental data integration, and learning from user interactions over time to weigh information and emphasize the relevant clinical, well-being and social determinants of health for a patient. A key step in the development of a care plan is to obtain a comprehensive holistic multi-faceted clinical, behavioral and social view of a patient. Collecting and refining the data from which a care plan is constructed in a timely manner is essential for the care of high cost/high needs individuals, who often deal with multiple chronic conditions and social issues [23]. Omission of relevant information can result in inefficient care and a failure to fully address the needs of the individual, beyond clinical issues. Furthermore, organizations also typically record relevant medical and social histories across several systems, notes and files.

The system proposed in this paper captures knowledge from care professionals' observations, often in an unstructured form, in order to create a holistic patient-centered view, consisting of entities extracted from free-text case notes. The system then provides care professionals with highlights on the most relevant information, based on its semantics, and informs them of the direction of care by displaying this information effectively. The key challenges in achieving these tasks are dealing with the diversity of the domain (e.g. social & clinical issues), the difficulty in building a single model to capture this multi-domain information and the effort required to manually map annotation artefacts to entities with explicit semantics.

An evaluation was undertaken with domain experts to investigate the proposed system. We discuss the results and open problems in building a cognitive approach to transform multidisciplinary information into meaningful views for person-centered care. We also conducted preliminary interviews with domain experts to gather early feedback on the value and

978-1-5090-4881-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ICHI.2017.28

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.



Summary | Plan | Programs | Data | History | Team

Laurie Thompson
76 Years
1/1/1940
Female
Priority: Not Set

Actions

Address
265 White Pond,
Stormville,
12582

Goals
Diabetes Education ●●●●● Excellent 12/6/2016
Exercise Plan ●●●●○ Good 12/6/2016
Medication Adherence ●●●○○ Satisfactory 12/6/2016

Actions
No Records

Assessments
No Records

Barriers
Transport
Mobility

Notes Highlights
Last Year
All
Last updated on 12/14/2016, 7:10 AM
depression
541-878-4534
hypertension
metformin to 15mg/Daily

Figure 1. Cropped screenshot from Watson Care Manager.

usability of Note Highlights. Our long-term vision is to introduce analytics to consume these entities to provide actionable insights in the practice of care professionals.

II. Approach

The system architecture, its components and methodology are described in this section. The system first annotates raw case notes (associated to each patient) in free-form text to extract entities, which are then connected to well defined clinical, well-being and social specific entities from predefined ontology models or domain vocabularies/ taxonomies, described later. The semantically matched entities can then be fed to a care management or another health platform. Figure 1 shows this functionality in the context of IBM Watson Care Manager (WCM) [12]. The user gets a view of the key information in the notes pertaining to a patient, with the ability to filter by category and by time. This is important as, across healthcare settings, direct interactions between staff are not sufficient to advance complex patient care. Multidisciplinary teams need access to shared, easily retrievable and concisely presented visual information. Given the number and complexity of situations and/or interactions a patient may experience it is recommended that the design of digital patient case note and annotated systems incorporate hyperfunctional navigation to (1) reduce cognitive overload for healthcare professionals evaluating patient care and (2) improve synthesis for the collaboration and exchange of critical information between healthcare professionals during the process of patient care. The overall impact is to both enhance awareness of the patient needs and correctly coordinate their care [4]. To support the above functionality, the following components have been implemented as part of the architecture shown in Figure 2: (a) the annotation component consolidates the annotations from different text annotators; (b) the terminology service maps the health and social care entities extracted from the annotated text to well-defined vocabulary entries and assigns them a semantic type; (c) the Note Highlights ranking component (illustrated in the circular flow) ranks and personalizes the information based on user feedback. This paper focuses on the annotation component and the terminology service, sketching the note highlights component.

A. *Annotators Façade*

Off-the-shelf annotators are used to extract the relevant pieces of information from text (Named Entities and annotations), such as case notes from care professionals or medical records. Specifically, we are using Advanced Concept Insights (ACI), an Unstructured Information Management Architecture (UIMA)-based IBM annotator for clinical text based on the 10th revision of the International Statistical Classification of Diseases (ICD-10) terminology [13] and AlchemyAPI for keywords and entities [1] to annotate concepts not included in ICD-10 – mainly social determinants. New annotators can be added using the Annotators Facade, which acts as a single-entry point, handling requests by routing to the appropriate service(s) and combining the results. As a proof point, for the second part of our validation, we added another UIMA IBM annotator, based on UMLS, which is integrated in the Electronic Medical Record Analysis system (EMRA) [7]. This paper will refer to this annotator as EMRA.

Annotations are filtered based on a minimum confidence threshold and are assigned a type from a pre-selected set of relevant types, called *View Types* (described in the Terminology section). The mapping between annotations and *View Types* is first done through a set of rules, varying per annotator: ACI

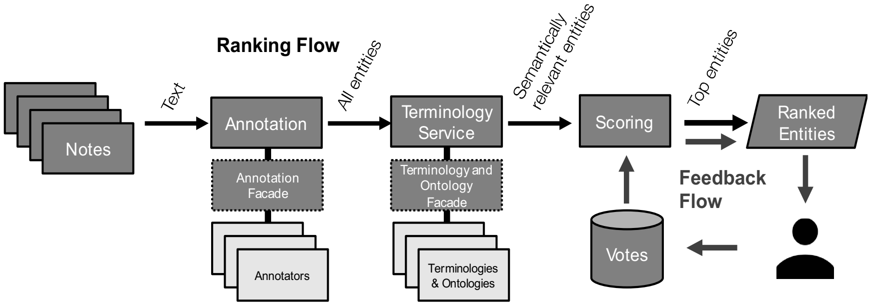


Figure 2. Web Service-based architecture.

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

retrieves an alphanumeric ICD-10 code for the annotations, which encodes the category of the concept in the ICD-10 hierarchy in an unambiguous way. Alchemy Entities may assign a type to the entities based on its own type hierarchy, or it can also retrieve DBpedia [3] or Freebase [8] identifiers. DBpedia is largely equivalent to Wikipedia for data, as a large knowledge base of open domain facts, and Freebase is also a large graph of open domain facts, curated collaboratively. A meaningful mapping is provided from ICD-10 codes and Alchemy types to *View Types*. In the case of EMRA, we return all UMLS annotations that a have semantic type that corresponds to one of our relevant *View Types*. We use the UMLS semantic hierarchy [11] to filter the relevant types. See the correspondence in Table II. This is not enough to assign types to all annotations, for the remaining ones, a *View Type* is assigned using the Terminology Section described in the following section.

Results from all annotators for a given note are combined. If different annotated entities are extracted by different annotators, for the same or overlapping text (i.e., they have the same start index in the case note), a reduction strategy is used to keep the most accurate annotation based on the following: (1) *View Type*, annotated entities with a known type are preferred over unknown types (2) the start and length for the covered text, longer annotations are preferred (example: “diabetes type 2” vs. “diabetes”) (3) the priority of the annotator, if provided (optional configuration). Thus, for a given start index in the original text, only one annotation entity is saved, the one with the highest ranking in the lexical order described above.

B. *Terminology Service*

The aim behind the Terminology Service is to map the annotated entities into a common well-defined vocabulary (based on a defined set of models), and assign them a *View Type*. Each *View Type* is assigned to a category as shown in Table 1. These types and categories have been defined through expert consultation. A *View Type* is assigned based on the semantic taxonomies of an entity, given by the relevant terminologies. The terminology is currently built by creating a Lucene inverted index [18], containing the entities URIs, preferable label, alternative labels (synonyms), and semantic types from the following ontologies: (1) The ICD-10 hierarchy of clinical terms, covering a broad range of conditions; (2) a Linked Data subset of Freebase covering the clinical domain, in particular, all the entities, which Freebase types (symptoms, risk factors, conditions, medication, treatments, procedures and medical specialties) have a clear correspondence with our *View Types*; (3) a Linked Data subset of DBpedia covering the *View Types* in Table 1.

TABLE I. CORRESPONDENCE OF UMLS TYPES TO VIEW TYPES.

UMLS Semantic network	**View Type**
CHEM|Chemicals & Drugs.	Medications
ACTI|Activities & Behaviors. LIVB|Living Beings| Age Group, Family Group, Professional or Occupational Group.	Social
CONC|Concepts & Ideas| Idea or Concept, Intellectual Product, Qualitative Concept, Quantitative Concept.	
DEVI|Devices. PROC|Procedures.	Procedures
DISO|Disorders.	Symptoms and Diseases
OCCU|Occupations	Services
ORGA|Organizations	Organisations
GEOG|Geographic Areas	Places
ANAT|Anatomy. GENE|Genes & Molecular Sequences. PHEN|Phenomena	No relevant *View Type*

While the clinical types are well defined in ICD10 and Freebase hierarchies, the social ones are not defined as such. Thus, we use DBpedia as a complement to Freebase. DBpedia entities are categorized following the Wikipedia categories, which have a good coverage of both clinical and social topics. Differently from Freebase, DBpedia categories are not flat but they are organized in a hierarchy. The *View Types* are manually matched to top level categories in the hierarchy, and entities are extracted using inference over the subsumption hierarchy. The DBpedia taxonomy is materialized up to a depth of three to extract all relevant entities in the subcategories. Through deductive inference, we increase the coverage of the models, e.g., the DBpedia category “Self Care” linked to the *View Type* Activities of Daily Living (ADL), comprises entities like “nutrition”, “sleep hygiene”, etc., as well as relevant subcategories such as “Physical exercise”, “Hygiene” and “Positive Mental attitude”. Unfortunately, this process also introduces noise, thus, the depth was set to three. DBpedia entities are linked to Freebase. Therefore, in order to use a unique URI (and limit duplication and irrelevant entities), we created a model with only the DBpedia entities that have a Freebase URI and belong to a semantic type up to the parent category that corresponds to our *View Type*, including all alternative labels, redirects, equivalent entities, and its original semantic category.

Given an annotated entity, the terminology service retrieves an indexed entity by performing a fuzzy text search over all the labels to find the entity matches. An indexed entity will have a unique URI from one of the models (an ICD-10 or a freebase URI), producing a consistent set of semantic entities independently of the source of the annotation (the annotators used by the system). The highest ranked entity is retrieved using a combination of Lucene full-text search and string distance metrics [5] to lexically rank the matches, and the expected type of the entity (if given by the annotators). If an entity has equally good matches in both models, ICD-10 has preference, therefore an ICD-10 based entity will be retrieved.

The added value behind using a terminology service is that having a unique URI gives a well-defined global meaning to handle heterogeneous annotations. In other words, annotations representing synonyms with alternative names but similar meanings will be represented with the same entity, providing a

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

shared, non-ambiguous, non-duplicated vocabulary. As an example, an entity consists of a unique identifier from Freebase (m.04v6hz), a pref. label (e.g., "Eyedrop"), alternative labels (e.g., "Eye-drops" "Ocular lubricant", etc.) and a *View Type* (e.g., "Procedure"). In addition to the terminology service finding types when the annotators fail to (using the static rules described above), it also merges different entities. For example, "eye drops" and "eye drop" are merged as one. The type given by the annotators, when known, is used to disambiguate when more than one representation from the terminology exists. We are investigating the addition of other annotators specialized in extracting findings from EMRs[27], which requires consolidating annotations from other taxonomies (in this case SNOMED CT).

If the models do not cover the annotated entity, the entity is assigned the *View Type* "unknown". As social entities are notably hard to identify, we do not discard unknown entities at this stage. We evaluate both the coverage and impact on precision and recall on keeping vs. filtering unknown entities.

C. Note Highlights Ranking.

View Types are not mutually exclusive (obesity can be both a risk factor or a symptom). These *View Types* have a broad coverage and the purpose is not to replace existent clinical or social hierarchies. The rationale behind using them is: (1) to present the information to the user based on a small set of clinical and social dimensions configured in advance (given by a client), and that abstracts from the annotators models of choice but are easily mapped to other clinical or social hierarchies; (2) To filter the information that is most likely to be relevant for the user based on the type (eliminating noisy or non-relevant entities) and to drive relevant recommendations, i.e., the most influential types that are used as the "features" from the learning algorithms (see Section 6 on long term vision). As an example, entities about contacts or home addresses are not "actionable" types to provide recommendations, but they may be relevant for a certain care worker dealing with the patient.

All of this context is very important, but it could lead to information overload for care professionals. Most of the information is unstructured, and while semantics can help organizing and linking entities into views, there is not a unique model of everything when mixing clinical with social information. The WCM interface shows Note Highlights to the user (Figure 1), all available entities are displayed to the care professional, organized according to a semantic view (based on the *View Types*) and a temporal view (last month, last 6 months, etc.), along with the supporting evidence if requested (e.g., the case notes where the entity was found). Ranking is needed to create Note Highlights with the most relevant entities. To do that the system needs to learn what is important for a specific patient profile, as well as what information is important for some users but not for others. WCM interface enables users in a care team to provide feedback in the form of up-votes (relevant entity) and down-votes (non-relevant entity). We have developed a method based on an adaptation of the Wilson score interval [28], which uses up- and down-votes to learn ranking of different entities according to three dimensions: user (what entities are usually important for this specific user), role (what entities are usually important for users having this role, e.g., a registered dietician vs. a primary care manager), and patient (what entities are usually important for all users – irrespective of their role – dealing with this patient). The details of this method and its evaluation are beyond the scope of this paper.

III. Evaluation and Results

A. Evaluation Setup and Metrics

We created a gold standard to measure whether our system can extract similar entities from notes as to what users, with mixed levels of expertise with the care domain, will chose. In other words, if the system is able to find all relevant and only relevant information. To create this gold standard, we needed to define: (1) the set of cases notes; (2) the set of relevance judgements (or the relevant entities for each case); (3) the evaluation metrics.

Set of cases. We selected 20 cases containing 36 existent patient notes from four different sources. Each case corresponds to an individual patient, contains between 1- 4 notes, and was evaluated by 4 experts, resulting in a total of 144 annotated notes. Each case has an average length of 492 words and the largest case has 1316 words. The cases were obtained from collections of notes, used for learning and training, and real care workers' notes. The cases were coded from A-T as this helped when keeping track of the cases assigned to evaluators and when the results of the evaluations were analyzed, as follows:

- Cases A-H: 8 clinical cases from the MT Samples [21] public collection of transcribed medical reports.
- Cases I-P: 8 care management cases extracted from a Care Management system. These are real cases that have been manually anonymized and de-identified.
- Cases Q-R: 2 clinical and social sample cases provided by Medicaid [10], as examples to narratively illustrate individual consumer's strengths and service needs.
- Cases S-T: 2 social care cases, based on real notes, used as illustrative examples to design personas for IBM products.

Table III shows descriptive statistics pertaining to the set of cases that were reviewed by domain experts. The shortest case (L) contained 14 distinct entities (including those extracted by the system or highlighted by 4 domain experts that acted as evaluators) and the longest case (A) contained 250 unique entities. The average number of entities for each case was 83 with a standard deviation (SD) of 56. The total number of entities that experts highlighted across all cases was 4469; 1656 entities were unique, and this contrasts with the number of entities identified by the system (2711).

TABLE II. DATASET CHARACTERISTICS.

Unique cases	20
Average case length (in words)	492
Unique notes	36
Annotated notes	144 (36 * 4)
Total unique user entities	1656
Total unique entities (system + user)	2098
Total user entities	4469
Total system entities	2711

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

The style of the notes differs significantly according to their provenance and this allows us to compare the performance of our methods using a diverse set of notes, across both clinical and social cases. All cases are different yet they have in common that the patients were previously diagnosed with diabetes and had also other conditions. The rationale behind this is twofold. On one hand, diabetes is one of the three most common chronic conditions in multimorbidity patients[24], requiring the implementation of lifestyle changes, thus making diabetic patients the target population in many structured care programs. On the other hand, it is more likely that the cases will share some of the entities, this will be useful to reuse this gold standard for a future evaluation on learning algorithms to predict relevant entities based on patient history similarities.

Relevance Judgements. To create the relevance judgements the evaluators were asked to: (1) read the notes pertaining to a single patient (a case) and highlight all the keywords that they consider relevant (user based annotations), and (2) rank the top-10 annotations (highlights) for a case.

We performed a preliminary evaluation with non-experts and a second evaluation with domain experts. The preliminary evaluation was used to test the feasibility of our validation approach and obtain preliminary results on the system performance. In here, 15 non-expert evaluators were recruited. These were all IBM employees who were not involved in the project and with different roles and different levels of domain expertise. The second evaluation provided the results from domain experts from two institutions: (1) Orlando Health, a Florida-based private, not-for-profit network of community and specialty hospitals; and (2) Trinity School of Nursing and Midwifery in Dublin, Ireland. In an on-site visit to Orlando Health, we recruited 14 evaluators with different roles: care coordinators - CC (mostly registered nurses - RN), care transition navigators - CTN, CC supervisor, CTN supervisor, quality data coordinator, population health coordinator, and RN risk coder, and whose experience in the care domain ranged from 2 months to 40 years (average 9.6, SD 9.7). In a parallel on-site visit to Trinity we recruited 6 additional evaluators, all nurses with different roles, including midwifery, women's health, pediatrics, radiology, surgical, older adults and oncology, and whose domain experience ranged from 3 to 33 years (average 17.8, SD 11.6). Therefore, we recruited a total of 20 nurses and care workers with hands-on extensive experience on the delivery of care.

Evaluators were explained that they need to select keywords (or key phrases) that capture not just clinical, but also social aspects that they consider relevant about a patient. A mock case note was given so as to familiarise the evaluators with the task. Evaluators could highlight the same keyword several times if they consider it important. They were also made aware that a keyword may consist of a combination of words if these are required to understand the meaning or importance of the passage. For example, "Eye drooping" was underlined as it would not be useful to just underline "eye".

For each highlighted annotation from the text, evaluators were also asked to assign a category. Requiring evaluators to select a category from all our *View Types* would have been a burdensome task. Therefore, to reduce the time evaluators take in each note, we simplified the task by just making the evaluators choose between the top categories as seen in Table I, namely Clinical, Services, Social, Places, Contacts and Other. Evaluators were given a detailed version of Table I that explained each category and that also stressed that this information should not influence their choice of relevant entities.

The explanation sessions with each evaluator lasted no longer than 20 minutes. Domain expert evaluators took on average 11 minutes to evaluate a case, 20 minutes the longest (non-expert evaluators took on average 15 minutes to evaluate a case and 27 minutes the longest). Each case was evaluated by 3 non-expert evaluators in the first run of the evaluation and by 4 different domain expert evaluators in the second run in order to determine user agreement.

The system is aimed at care team members with interdisciplinary expertise who may not necessarily be clinical experts (e.g. care workers, community providers, informal care givers, nurses, etc.). Social aspects are not as well defined as clinical aspects, and while we have replicated this evaluation with both non-domain and domain experts with different backgrounds, for the domain expert evaluation the social and clinical annotator EMRA was added to the annotators façade (in addition to ACI and Alchemy). Except for the addition of EMRA, the exact same system was used for both the internal and domain expert evaluation.

With regards to relevance judgements, these were obtained from the text that was highlighted by the evaluators. The way evaluators highlight (i.e., underlined) words, sentences and concepts can vary significantly and, in turn, this may affect the identification and selection of relevant entities. This is particularly important as entities selected by the evaluators need to be matched with those identified by the annotators. In order to mitigate the aforementioned risks, a systematic approach was used to obtain the relevant user annotated entities for the purposes of this evaluation. The approach relies on a first removal of stop words (e.g. "who", "the", "that", "is") from the passages highlighted by the evaluators. Later, entities are split when commas, conjunctions or prepositions appear, provided that this splitting does not change the meaning of the individual entities. Entities that are clearly different are also separated. For instance, "33-year-old female" consists of two entities: "33-year old" and "female". Duplications and plurals are also either removed or merged with their corresponding root entities. Finally, contextual modifiers are taken into account and separated from the entities when the meaning of the passage remains unaltered. We identified four different types of contextual ambiguity:

- Temporal ambiguity, such as dates, times or expressions including "past", "tomorrow", or "long-term"
- Measurements, such as drug posology or laboratory test results (e.g. "Hemoglobin of 14 g/dl")
- Qualifiers, such as "mild", "severe", or "large"
- Negations, such as "not well" or "denies" (e.g., "denies joint pain".

When applicable, contextual modifiers were marked and every action was logged. This systematic approach includes a clerical review, where three trained staff members transcribed,

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

assessed and cross-validated each other's work. This was a time-consuming effort but one that allowed us to compare, with confidence, those entities selected by evaluators with those identified by the annotators.

In turn, this has allowed us to measure precision and recall for the annotators independently of their abilities to capture the modifiers attached to the entity. In future work, we will evaluate the impact of such modifiers and whether users consider them to be relevant. Strategies based on linguistic pre/post-processing could be implemented to configure the relevant context that should be presented and associated to each entity. This could be further personalized according to the semantic type of entities and / or the role of the user (e.g., show drug posology of patients to health professionals who write prescriptions).

Evaluation metrics. We measured inter-rater agreement in the following way: Strong agreement if three or more evaluators pick up the same annotation; Moderate agreement if two evaluators pick up the same annotation; Weak agreement if only one. We observe an average score of .63, .19 and .17 for strong, moderate and weak agreements respectively across all domain expert evaluators. Agreement was stronger for experts than non-experts (non-expert evaluators had an average agreement of 0.5, 0.3 and 0.2 for strong, moderate and weak), which was to be expected given that they would be more familiar with the content in the notes. Regardless of the diverse set of notes, none of the experts commented about intelligibility or comprehensibility of the notes. The results are listed in Table IV. We consider an annotation to be relevant if there is moderate or strong agreement across users (two or more evaluators highlighted it). Precision (P) and Recall (R) was calculated in the manner described in Figure 3 for all entities extracted by the system with respect to all relevant user annotations. P measures the system ability to find accurate annotations (not noisy). R measures the system ability to find all relevant annotations for the evaluators (coverage). The F_1 score is the weighted average of P and R and provides a measure of how balanced these two metrics are.

B. Results.

Results for domain experts are presented in Table IV. For each case, we indicate:

- Case Provenance: we have 8 clinical cases (A-H from MTSamples), 2 mixed clinical and social cases (Q-R) and 10 social care cases (I-P, S-T), containing also health information.
- User agreement: strong (S), moderate (M) or weak (W).
- Total number of entities per case (N Entities): as annotated by evaluators.
- Average time evaluators took for each case. Average P, R and F_1 considering all annotations returned by the system.

$$\mathrm{P} = \frac{\text{System relevant entities}}{\text{System annotations}}$$

$$\mathrm{R} = \frac{\text{System relevant entities}}{\text{User relevant annotations}}$$

$$\mathrm{F1} = \frac{2\,(\mathrm{P} \times \mathrm{R})}{\mathrm{P} + \mathrm{R}}$$

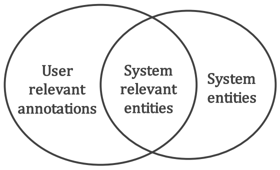


Figure 3. Evaluation Metrics.

TABLE III. DOMAIN EXPERT VALIDATION RESULTS.

	Agreement %	N	Time	All Annotations		
Case	**S / M / W**	**Entities**	**(min)**	**P**	**R**	**F1**
A	.46/.24/.28	200	20	.56	.90	.69
B	.85/.09/.05	75	8	.75	.92	.83
C	.52/.36/.10	137	10	.69	.90	.78
D	.67/.17/.14	184	20	.75	.83	.79
E	.53/.25/.21	107	12	.6	.85	.70
F	.47/.25/.27	92	11	.64	.82	.72
G	.72/.19/.07	66	14	.65	.83	.73
H	.63/.17/.19	139	13	.7	.93	.80
I	.43/.25/.31	32	9	.47	.77	.58
J	.56/.18/.25	16	3	.64	.91	.75
K	.75/.15/.10	20	6	.77	.94	.85
L	.88/.11/0	9	5	.63	.77	.7
M	.41/.27/.30	79	14	.40	.78	.5
N	.47/.36/.16	125	18	.52	.81	.64
O	.76/.05/.17	17	3	.72	.92	.81
P	.85/.04/.09	21	4	.77	.89	.82
Q	.73/.16/.10	67	11	.60	.8	.69
R	.60/.14/.25	70	14	.66	.76	.71
S	.54/.21/.24	85	8	.55	.92	.69
T	.72/.20/.07	115	18	.81	.83	.82
Average	**.63/.19/.17**	**83**	**11**	**.64**	**.85**	**.73**
Avg. A-H	.61/.21/.16	125	14	.67	.87	.75
Avg. I-P	.64/.18/.17	40	8	.62	.85	.71
Avg. Q-R	.66/.15/.18	69	12	.63	.78	.70
Avg. S-T	.63/.20/.16	100	13	.68	.87	.75

In Table V we compare the results of the internal validation with those from the domain experts' validation. This comparison includes only two of the annotators (ACI and Alchemy) as EMRA was not available at the time the internal validation was undertaken. Table V shows that domain experts were faster and more often in agreement with each other than the non-experts. Overall the internal validation showed larger P and R. This could be because first domain experts highlighted less entities as relevant than the non-expert internal evaluators, and those entities were found by the system (therefore affecting precision); and second the internal evaluators were more familiar overall with the concept of "an entity" and highlighted less complex entities that are hard to catch by annotators than domain experts (therefore, affecting recall, as for example: "eats on the run" "does not want to live anymore").

In addition, we aimed to answer the following questions:

- What is the effect on P/R when several different metrics are considered (Table VI)?

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

- What is the effect on P/R when looking at each annotator individually (Table VI)?
- What is the system coverage on Top-10 user annotations (highlights) and those with strong agreement (Table VII)?

TABLE IV. COMPARISON OF INTERNAL AND EXTERNAL VALIDATIONS.

	Agreement		Time	All Annotations		
Case	**S / M / W**	**N Entities**	**(min)**	**P**	**R**	**F1**
Domain expert validation results (only ACI & Alchemy based)						
Avg. All	.63/.19/.17	83	11	.78	.48	.59
Internal validation results (only ACI & Alchemy based)						
Avg. All	.5/.3/.2	88	15	.86	.56	.67
Avg. A-H	.5/.3/.2	130	18	.86	.58	.69
Avg. I-P	.6/.2/.2	44	11	.88	.53	.65
Avg. Q-R	.4/.4/.2	58	12	.78	.51	.61
Avg. S-T	.7/.2/.1	119	17	.86	.66	.74

Figure 4 shows the effect of different metrics (e.g. only considering strong user agreement, excluding noise, etc.) on P/R compared to all annotations. When looking at strong user agreement compared to all annotations, recall increased to .89. This is an indication that the system has potentially good coverage for the entities that matter the most, and that with enough user training to rank the relevant entities, the F1 of the system can increase over time. When only including known *View Type, there was a* small negative effect in R, from .85 to .74 for known types. This is because several entities do not have a *View Type* due to a lack of model coverage for some types (e.g., "bedtime snack" was given an unknown type as it couldn't be identified as an *Activity of Daily Living*). However, precision increases slightly. There is a trade-off between precision and recall and the system can be configured to favour precision by choosing only entities with known type if required.

The reason we see precision being less than recall is overall due to two key factors:

(1) weak user agreement; If any user agreement is included (at least 1user picked entity) then precision increases from .64 (all annotations) to .78, and

(2) noise picked by annotators; noise was included in all annotations and this affects precision because 21% of the entities in each case were noise. When noise is excluded, precision increases to .85. Strategies based on linguistic processing can be used to filter out some of the noise, system annotations like "will" or "Spoke" with the type "person" can be removed if they have "VERB" as "Part of Speech" (POS).

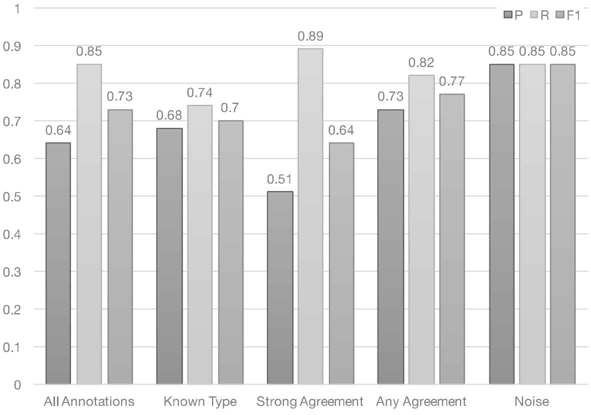


Figure 4. Comparison of different metrics and their effect on Precision, Recall and F1.

The effect on P/R when considering EMRA, Alchemy or ACI separately was also considered. Compared to ACI and Alchemy, EMRA's precision increased the least (.64 to .66). This is due to the amount of noise that EMRA introduced. Conversely, ACI had the highest precision increase (.64 to .81) as well as the largest decrease in recall (.85 to .3). These results show the importance of including social concepts beyond ICD-10 clinical terminology to greatly improve the recall.

In Table VII, we examined the system coverage on the top ranked entities by the experts (Top-10) where we look at the coverage (i.e., recall) of the system for the experts: (1) the top-10 entities selected by expert evaluators (at least 1 evaluator chose as their top entity); and (2) the top-10 entities selected by at least two evaluators. The proportion of top entities picked up by the system was 83% and this increased to 91% when considering strong agreement.

TABLE V. EFFECT OF DIFFERENT METRICS ON PECISION, RECALL AND F1

Cases	Known type		Strong Agreement (>2 users)		Any Agreement (=>1 user)		Excluding Noise		EMRA		Alchemy		ACI	
	P / R	**F1**	**P / R**	**F1**	**P / R**	**F1**	**P / R**	**F1**	**P / R**	**F1**	**P / R**	**F1**	**P / R**	**F1**
A-H	.70/.79	.74	.51/.89	.64	.78/.84	.81	.85/.87	.86	.67/.80	.73	.84/.25	.37	.86/.37	.50
I-P	.65/.71	.67	.52/.90	.64	.71/.81	.76	.86/.85	.85	.65/.80	.71	.83/.38	.50	.77/.27	.40
Q-R	.63/.60	.61	.51/.78	.62	.79/.79	.79	.80/.78	.79	.63/.69	.66	.43/.17	.25	.69/.15	.24
S-T	.72/.78	.74	.54/.91	.66	.56/.82	.76	.88/.87	.87	.68/.79	.72	.79/.49	.60	.83/.34	.49
All	.68/.74	.70	.51/.89	.64	.73/.82	.77	.85/.85	.85	.66/.79	.71	.79/.49	.60	.81/.30	.43

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

TABLE VI. PROPORTION OF ENTITIES SELECTED BY EXPERTS AS THEIR TOP-10 ENTITIES THAT WERE ALSO RETRIEVED BY THE ANNOTATORS.

Cases	Top-10 (%)	Top-10 Strong Agreement (%)
A-H	85	94
I-P	81	87
Q-R	71	85
S-T	82	90
All	**83**	**91**

Finally, we investigated precision and recall for the entity types by comparing the types assigned by the system for all the relevant entities to the types assigned by experts. We defined that a system type is accurate for an entity if at least one user has assigned the same type to that entity. Note that some entities could have an ambiguous type and different users may choose different types, for example "brother" is marked both as "social (family)" and as "contact". Thus, for all relevant entities, precision is calculated as all accurate system types with respect to all known types assigned by the system; while recall corresponds to all accurate system types with respect to all known and unknown types given by the system. Results on P/R for types are shown in Table VIII, together with how many entities were classified by users as clinical (CL), social (SO), contacts(CO), services(SE), places(PL), or any other (OT) type. Cases A-H are predominantly clinical case notes and this is reflected in the number of CL types seen in Table VIII (80.1). Table VIII also includes the average number of unknown entities with respect to the average number of unique entities, as well as the precision (i.e., entities with an accurate type types with respect to all entities with a type) and recall (i.e., entities with an accurate type with respect to all entities, with or without a type assigned by the system).

TABLE VII. P/R FOR SYSTEM TYPES COMPARED TO USER TYPES.

Cases	Types given by users						N^{O}unk/	P	R	F1
(Avg)	CL	SO	SE	CO	PL	OT	N^{O}ent			
A-H	80.1	15.6	6	2.2	2.5	1.7	6/88.7	.8	.8	.8
I-P	18.2	7.4	3.7	3.7	.7	3.5	3.5/26.3	.8	.6	.7
Q-R	33.5	22.5	.5	5	.5	0	8.5/45.5	.8	.6	.7
S-T	47.5	29	5.5	9.5	2	2.5	7.5/73.5	.8	.7	.8
All	**47.4**	**14.3**	**4.5**	**3.8**	**1.5**	**2.3**	**5.4/57.9**	**.8**	**.7**	**.7**

As shown in Table VIII, only 9.3% of entities (5.4/57.9) could not be assigned a type, with an F1 of 0.7.

IV. Discussion

In this section, we discuss why annotators failed to identify some of the relevant annotations indicated by the evaluators (recall) and discuss the effectiveness of the system. First, clinical notes have slightly better P/R. This was expected as models currently have better coverage of clinical entities than social ones. Apart from a lack of models' coverage and unknown organisation-specific acronyms, annotation granularity is the second reason user annotations were missed. In other words, evaluators highlighted as relevant not just entities but also the context surrounding them, which in the following cases was not picked up by the annotators:

- Factual changes and actions, whether they are clinical (e.g., "gained weight", "changed her medication", "left side of her face is dropping", "stop taking the insulin") or social (e.g., "managing her husband health", "lost his job", "achieve that goal").
- Some complex entities (noun phrases with more than one noun) (e.g., "lost 9lbs", "lives with mother", "calories below 1000", "eat 3 meals a day", "inflated self worth"). Nonetheless it can find lab and medication measures such as "blood sugars in the 20's" and "80mg of metformin once/day".
- Feelings and emotional status (e.g., "energetic", "feels very alone", "overwhelmed", "lost all interest", "cried for two minutes") and marital status such as "married" "widowed"

Structured 'facts' are different from entities. For example, consider the text "John quit smoking for several years till he picked up the habit recently". The annotators extract the entities: "quit smoking for several years", "habit" and "smoking", failing to pick up the true meaning behind this sentence. Solving this challenge requires advances in natural language processing to extract facts from text, and perhaps more complex semantic models to structure the patient-centred information, negations vs. missing information and temporal relations into a patient profile.

Summarising the results of the previous section, we are making the following key observation: agreement was stronger for experts than non-experts, expert users showed different behaviour when annotating text, being more selective in what they considered relevant as well as being faster in completing the task. This highlights the fact that expert input can differ significantly from non-expert input, even for a relatively simple task. In light of this, studies need to consider expert input to ensure validity. For the validation with experts, the system was very effective at identifying relevant entities (85% recall) and only those entities (64% precision). Removing noisy annotations resulted in the same recall but increase precision (85%). In addition, the system was very effective at identifying the entities that were considered most important by the experts. For any of these entities, coverage was 83% for entities marked as important by any user and 91% for entities for which there was strong agreement among experts regarding their importance. The above illustrates that the system is a feasible solution in identifying entities that are relevant for healthcare professionals.

V. Feedback from domain experts

In addition to evaluating our system quantitatively, we also interviewed six experts - Care Managers / Coordinators (CC), which are registered Nurses, and Care Transition Navigators (CTN) from two US-based healthcare organisations. The purpose of these individual hour-long interviews was to conduct a preliminary design review to assess the value of Note Highlights to domain experts in Care Management and to identify potential issues and challenges. We report here the

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

feedback and insights on potential usability and desirability, in terms of key features and workflows rather than UI design. For future work we plan to measure usefulness, based on whether the system can improve the way users share information or save them time when performing certain daily activities (e.g., when reviewing the patient information before calling them) based on displaying patient information effectively, as well as validating if Note Highlights can be trained over time, while in use, to improve its efficiency, therefore, measuring the impact of this learning in a realistic scenario.

A. *Interview Structure*

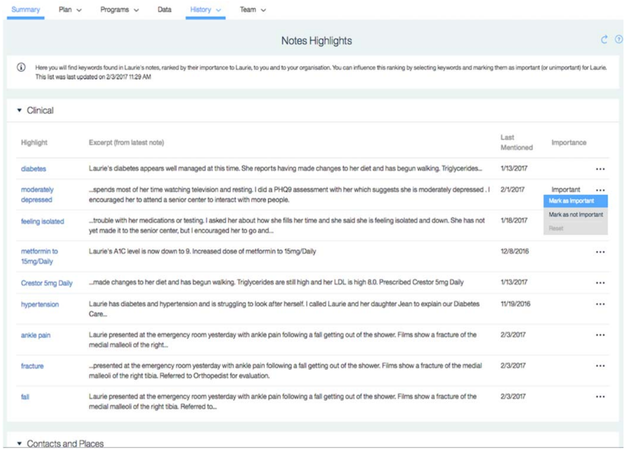
Initially, the experts were asked about how they prepare for an interaction with a care-managed patient, focusing on their use of unstructured case notes. They were presented with a simple patient scenario and asked to use our system to prepare (gather information) for their next call with that patient. In the scenario, they were told some important details of the patient's history. Other details were intentionally omitted, for example, that the patient had a recent fall and trip to the emergency room. The experts were invited to use our prototype, without any training, to prepare for an interaction with this patient. They were asked to ‘think out loud’ about what they were doing and thinking. If they became stuck or unsure, then the aspect of the system they found unclear was explained. This enabled both the ’first use’ user experience and the ongoing value of the system to be assessed. As part of this patient scenario, the system presented the experts with a ranked list of keywords (“Note Highlights”). Users could click on a highlight to read all case notes containing that word or phrase. They could also optionally mark a highlight as important, which caused the highlight to be prominently displayed for that user, as shown in Figure 5. The feedback received also enabled the system to learn which words matter most to particular patients and user roles, so that highlights could be ranked and presented in a personalised way. After a scenario briefing, the experts were asked to use our system to review case notes in preparation for their next call with the patient.

B. *Interview feedback:*

Use of unstructured case notes. All experts reported that it was essential to know the patient well before any interaction. They see several patients and need to refresh their memory of a patient or catch up on patient's current situation.

The experts reported investing anywhere between 5 to 30 minutes preparing for each patient. They considered case note review to be an essential, albeit time consuming, part of their work. Notes about interactions with other care professionals (physician, dietician, etc.) were of particular interest. Some illustrative quotes included: “I read notes before every call, … I focus on the things that are important” (CTN). “I'm looking for what I don't know.” (CC).

Time Value. The experts generally made effective use of the highlights. They quickly spotted those highlights in the list that interested them and used them to directly access the notes they felt were most useful to them. Their feedback tended to focus on potential time-savings. Some illustrative quotes included:” I'd find this useful to catch up with a patient I haven't seen in a while.” (CC). “This is good, it gets to the point.” (CC). “I would find that beneficial. If I only had 5 mins before calling a patient, this would be beneficial”. (CTN)



Notes Highlights

Here you will find keywords found in Laurie's notes, ranked by their importance to Laurie, to you and to your organisation. You can influence this ranking by selecting keywords and marking them as important (or unimportant) for Laurie. This list was last updated on 2/3/2017 11:29 AM

▾ Clinical

Highlight	Excerpt (from latest note)	Last Mentioned	Importance
diabetes	Laurie's diabetes appears well managed at this time. She reports having made changes to her diet and has begun walking. Triglycerides...	1/13/2017	···
moderately depressed	...spends most of her time watching television and resting. I did a PHQ9 assessment with her which suggests she is moderately depressed . I encouraged her to attend a senior center to interact with more people.	2/1/2017	Important ···
feeling isolated	...trouble with her medications or testing. I asked her about how she fills her time and she said she is feeling isolated and down. She has not yet made it to the senior center, but I encouraged her to go and...	1/18/2017	
metformin to 15mg/Daily	Laurie's A1C level is now down to 9. Increased dose of metformin to 15mg/Daily	12/8/2016	···
Crestor 5mg Daily	...made changes to her diet and has begun walking. Triglycerides are still high and her LDL is high 8.0. Prescribed Crestor 5mg Daily	1/13/2017	···
hypertension	Laurie has diabetes and hypertension and is struggling to look after herself. I called Laurie and her daughter Jean to explain our Diabetes Care...	11/19/2016	···
ankle pain	Laurie presented at the emergency room yesterday with ankle pain following a fall getting out of the shower. Films show a fracture of the medial malleoli of the right...	2/3/2017	···
fracture	...presented at the emergency room yesterday with ankle pain following a fall getting out of the shower. Films show a fracture of the medial malleoli of the right tibia. Referred to Orthopedist for evaluation.	2/3/2017	···
fall	Laurie presented at the emergency room yesterday with ankle pain following a fall getting out of the shower. Films show a fracture of the medial malleoli of the right tibia. Referred to...	2/3/2017	···

▾ Contacts and Places

Figure 5. Full screenshot of Note Highlights.

Quality Experience. In their feedback, several experts also mentioned quality. The case notes contained information that was not communicated to the experts in the initial scenario description, i.e. a fall and emergency room visit. However, most of the experts immediately spotted the keyword “fall” in the highlights and successfully drilled in to access the related notes. “I see ‘fall’ - straight away, I want to know more about that - was she injured?” (CC). One expert also mentioned that occasionally a tired or distracted Care Manager might miss important information in the case notes. She felt that the system could add value in reducing this type of human error. The following challenges were also identified:

Effective Learning. The experts were enthusiastic about the system tailoring keyword rankings to their specific role, but less enthusiastic about training it. Some were willing to use the simple “mark as important” feedback mechanism to train the system, others were not. A more reliable and consistent mechanism is needed to support effective learning. For example, the system could learn directly about which keywords the users choose to interact with - their “digital exhaust”.

Noise. Occasionally, the system identifies a meaningless keyword (e.g.: “normal”). Several of these were deliberately introduced to assess their impact on experts' trust in the system. Rather than ignoring them or marking them as not important, most experts expressed surprise and inspected them closely, trying to understand why they showed up. This suggests that noisy keywords may negatively impact usage efficiency.

VI. Conclusions and Long Term Vision

Reviewing and trawling through case notes is time consuming. The approach presented in this paper aims to simplify and improve the efficiency of this process. The combination of off-the-shelf annotators and a terminology service, on top of text notes, provides us with a view of a patient comprised of semantic entities. This view is based on a shared terminology that reuses well-known heterogeneous ontological models. For care professionals, Note Highlights must be consumable in a more efficient manner than simply reading the

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.

raw notes. This evaluation gives a baseline on what results we can expect. The promising results show that the system is capable of extracting meaningful entities about a patient when compared to what users would have annotated and highlighted. For future evaluations, we expect to first measure the effect of training the system on what entities are most relevant according to domain experts, and second, to determine whether Note Highlights offer a better/more productive experience for a care worker than simply reading case notes (usability study). This is a first step towards our long-term vision plan to create a cognitive system able to take the consolidated annotations and entity views to generate suggestions and insights for a patient, i.e., new relevant entities associated with the known ones. To do that, the integrated patient summaries can be fed into recommendation and prediction methods that use mixed learning models, based both on historical patient data and supporting evidence from domain knowledge sources covering multiple aspects of health.

In sum, we believe cognitive technology can help us to provide more efficient care, leveraging existent structured and unstructured domain and patient-centred knowledge, as well as learning from experts to provide better actionable insights, adapted to the knowledge acquired by an organisation over time.

Acknowledgments

We'd like to thank the School of Nursing and Midwifery in Trinity College Dublin and Orlando Health for providing the subject matter expertise on which this research is based - Suzanne Gruszka, Christine Delgado, Kim Weaver, Robyn Bartek and all the Care Coordinators and Care Transition Navigators. We'd like to thank James Shannon, Yenny Blanco, Ciara Layden, Niamh Caprani, Charlyne Lefebvre-Paille, Denisa Moga, Caroline O'Connor, the Watson Care Manager, the IBM EMRA and ACI teams and the IBM Design team in Dublin for their contribution and feedback.

References

[1] Alchemy API: http://www.alchemyapi.com/

[2] Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., Guthrie, B., Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet, 380(9836), (2012), 37–43.

[3] Bizer, C., Heath, T., Berners-Lee, T. Linked data-the story so far. Semantic Services, Interoperability and Web Applications (2009), 205-227.

[4] Bringay S., Barry C., Charlet J., Annotations for the collaboration of healthcare professionals. AIMA Annual Symposium Proceedings (2006), 91-96

[5] Cohen, W., Ravikumar, P., Fienberg, S., E., A Comparison of string distance metrics for name-matching Tasks. In IJCAI Workshop on Information Integration, 2003.

[6] Computing, cognition and the future of knowing: How humans are machines are forging a new age of understanding. IBM white paper at: http://www.research.ibm.com/software/IBMResearch/multimedia/Computing_Cognition_WhitePaper.pdf

[7] Devarakonda M., Tsou C-H., Automated Problem List Generation from Electronic Medical Records in IBM Watson. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas, pp 3942-3947, January 2015.

[8] Freebase: https://developers.google.com/freebase/.

[9] Hoc JM. Towards a cognitive approach to human-machine cooperation in dynamic situations. International Journal of Human-Computer Studies. 2001;54:509–540.

[10] http://www.health.ny.gov/health_care/medicaid/publications/docs/olcm/oltclcm-1att2.pdf

[11] https://semanticnetwork.nlm.nih.gov/download/SemGroups.txt

[12] IBM WCM : https://www.ibm.com/watson/health/population-health-management/care-management

[13] ICD-10: http://apps.who.int/classifications/icd/en/

[14] IDC Health Insight. Perspective: 360–Degree view of patients — can It be done?. At: http://www.idc.com/getdoc.jsp?containerId=HI251774

[15] J. Brooke, SUS - A "quick and dirty" Usability Scale, in: P. Jordan, B. Thomas, B. Weerdmeester, A. McClelland (eds.), Usability Evaluation in Industry, Taylor and Francis, London, 1996.

[16] Laxmisan, A, McCoy, A.B., Wright, A., Sittig, D.F., Clinical summarization capabilities of commercially-available and internally-developed Electronic Health Records.Appl Clin Inform, 3(1), (2012), 80-93.

[17] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, et al., DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal, 6(2), (2015),167–195,

[18] Lucene: https://lucene.apache.org/

[19] Marmot, M., Wilkinson, R., Social determinants of health. Oxford University Press (2009)

[20] McAuliffe, C., Experiences of social workers within an interdisciplinary team in the intellectual disability sector. *Critical Social Thinking: Policy and Practice* 1, (2009), 125-143.

[21] MT Samples medical online collection: http://mtsamples.com/

[22] Pivovarov, R., Elhadad, N., Automated methods for the summarization of electronic health records. J Am Med Inform Assoc 22(5), (2015), 938-47.

[23] Rigby, M., Hill, P., Koch, S., Keeling, D., Social care informatics as an essential part of holistic health care: a call for action. Int. J of Medical Informatics, 80(8), (2011), 544-554.

[24] Royal College of General Practitioners. Responding to the needs of patients with multimorbidity. A vision for general practice (2016)

[25] Schiff, G. D., Bates, D. W., Can electronic clinical documentation help prevent diagnostic errors? New England J of Medicine, 362(12), (2010), 1066–1069.

[26] Stevenson, J.E., Nilsson, G., Nurses' perceptions of an electronic patient record from a patient safety perspective: a qualitative study. J Adv Nurs. 68(3), (2012), 667-76.

[27] Tsou C-H., Devarakonda, M., Liang J. J., Toward generating domain-specific / personalized problem lists from Electronic Medical Records. AAAI Fall Symposium Series, 2015.

[28] Wilson, E. B., Probable inference, the law of succession, and statistical inference. J American Statistical Association 22 (1927), 209–21.

Authorized licensed use limited to: IBM. Downloaded on September 27,2023 at 11:13:24 UTC from IEEE Xplore. Restrictions apply.