# Bridging the Gap between Structured and Unstructured Health-Care Data through Semantics and Sentics[*]

Erik Cambria
COSIPRA Lab
University of Stirling, UK
eca@cs.stir.ac.uk

Amir Hussain
COSIPRA Lab
University of Stirling, UK
ahu@cs.stir.ac.uk

Chris Eckl
Sitekit Labs
Sitekit Solutions, UK
chris.eckl@sitekit.net

## ABSTRACT

As Web 2.0 dramatically reduced the cost of reaching others, forming groups, obtaining and republishing information, today it is easy and rewarding for patients and carers to share their personal experiences with the health-care system. This social information, however, is often stored in natural language text and hence intrinsically unstructured, which makes comparison with the structured information supplied by health-care providers very difficult. To bridge the gap between these data, which though different at structure-level are similar at concept-level, we exploit the semantics and sentics, i.e. the cognitive and affective information, associated with on-line patient opinions and, hence, provide the end-users of the health system with a common framework to compare, validate and select their health-care providers.

## Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: User interface management systems—*Design*; I.2.1 [**Applications and Expert Systems**]: [Natural language interfaces]

## General Terms

Theory, Design

## 1. INTRODUCTION

In health-care, it has long been recognized that, although the health professional is the expert in diagnosing, offering help and giving support in managing a clinical condition, the patient is the expert in living with that condition. Health-care providers need to be validated by somebody outside the medical departments but, at the same time, inside the health-care system. The best candidate for this is not the doctor, the nurse or the therapist but the real end-user of health-care – none other than the patient him/herself.

Patient 2.0 is central to understanding the effectiveness and efficiency of services and how they can be improved. The patient is not just a consumer of the health-care system but a quality control manager – his/her opinions are not just reviews of a product/service but more like small donations of experience, digital gifts which, once given, can be shared, copied, moved around the world and directed to just the right people who can use them to improve health-care locally, regionally or nationally.

Web 2.0 dropped the cost of voice, of finding others 'like me', of forming groups, of obtaining and republishing information, to zero. As a result, it becomes easy and rewarding for patients and carers to share their personal experiences with the health-care system and to research conditions and treatments. To bridge the gap between this social information and the structured information supplied by health-care providers, we exploit the semantics and sentics, i.e. the cognitive and affective information, associated with patient opinions over the Web, and hence provide the real end-users of the health system with a common framework to compare, validate and select their health-care providers.

## 2. STRUCTURING THE UNSTRUCTURED

In order to give structure to on-line patient opinions, we extract both the semantics and sentics associated with these in a way that it is possible to map them to the fixed structure of health-care data. This kind of data, in fact, usually consists of ratings that associate a polarity value to specific features of health-care providers such as communication, food, parking, service, staff and timeliness. The polarity can be either a number in a fixed range or simply a flag (positive/negative).

We propose to add structure to unstructured data by building semantics and sentics on top of it (Fig. 1). In particular, given a textual resource containing a set of opinions $O$ about a set of topics $T$ with different polarity $p \in [-1, 1]$, we extract, for each $t \in T$, the subset of opinions $o \subseteq O$ concerning $t$ and determine $p$. In other words, since each opinion can regard more than one topic and the polarity values associated with each topic are often independent from each other, in order to perform the mapping we need to extract, from each opinion, a set of topics (Section 2.1) and then, from each topic detected, the polarity associated with it (Section 2.2). Since both the procedures work at semantic level, they can be combined in a unique process having opinions as input and both semantics and sentics as outputs (Section 2.3).

---

## 2.1 Extracting Semantics

The extraction of semantics associated with patient opinions exploits ConceptNet, a directed graph representation of common sense knowledge [7], CF-IOF (concept frequency - inverse opinion frequency), a statistical method for the identification of common semantics [1], and spectral association, a technique that expands semantics through spreading activation [8]. In particular, we apply CF-IOF on a set of 2000 topic-tagged posts extracted from Patient Opinion [11], a social enterprise providing an on-line feedback service for users of the UK National Health Service (NHS), and use spectral association in order to find domain dependent concepts, which are stored in a database to be accessed at run-time by the opinion analysis process.

Thanks to CF-IOF weights, it is possible to filter out common concepts and detect domain dependent concepts that individualize topics typically found in patient opinions such as cleanliness, food, kindness of staff and timeliness. These concepts represent seed concepts for spectral association, which spreads their values across the ConceptNet graph and, hence, detects semantically related concepts concerning the same topic.

## 2.2 Extracting Sentics

The extraction of sentics associated with patient opinions exploits AffectiveSpace, a language visualization and analysis system [2], the Hourglass of Emotions, a novel emotion categorization model [4], and a human emotion ontology (HEO). In particular, we merge ConceptNet and WordNet-Affect (WNA) [13], a linguistic resource for the lexical representation of affect, and apply singular value decomposition on the resulting matrix to obtain AffectiveSpace, a multi-dimensional vector space of affective common sense knowledge. This vector space is then organized, using a k-means clustering approach, with respect to the Hourglass model (i.e. by using the sentic levels as 'centroid concepts'), and used to infer the affective valence of concepts, in terms of Pleasantness, Attention, Sensitivity and Aptitude, according to their relative positions (i.e. their dot product) in the space. Thanks to this process, therefore, we can assign to any given concept a sentic vector, from which it is possible to calculate affective information such as emotions conveyed or polarity. The overall polarity of a patient opinion, in particular, is obtained by averaging out all the concepts' polarity values, according to the following formula:

$$ p = \sum_{i=1}^{N} \frac{Plsnt(o_i) + |Attnt(o_i)| - |Snstv(o_i)| + Aptit(o_i)}{9N} $$

where N is the total number of retrieved concepts and 9 is the normalization factor (the maximum value of the numerator in fact is given by the sentic vectors [3, $\pm$3, 0, 3] and the minimum by [−3, 0, $\pm$3, −3]). In the formula, Attention and Sensitivity are taken in absolute value since, from the point of view of polarity rather than affection, all of their sentic values represent positive and negative values respectively (e.g. 'anger' is positive in the sense of level of activation of Sensitivity but negative in terms of polarity and 'surprise' is negative in the sense of lack of Attention but positive from a polarity point of view).

Additionally, in order to represent the resulting affective information in a Semantic Web aware format, we encode sentics in RDF/XML using the descriptors defined by HEO.
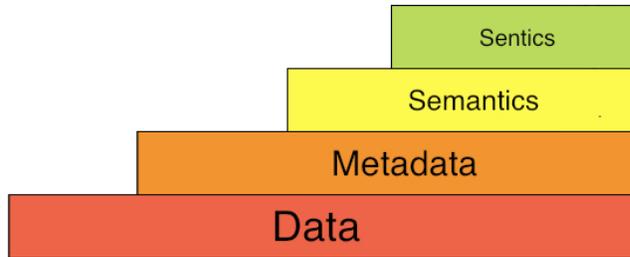


Figure 1: Semantics and sentics stack

This allows information to be stored in a Sesame triple-store, a purpose-built database for the storage and retrieval of RDF metadata [12]. Sesame can be embedded in applications and used to conduct a wide range of inferences on the information stored, based on RDFS and OWL type relations between data. In addition, it can also be used in a standalone server mode, much like a traditional database with multiple applications connecting to it.

## 2.3 Opinion Analysis Process

Patients, relatives and even health-care professionals have been telling their stories in support forums and on personal sites since the Web began but the information they have been providing is narrative, loosely structured and - like most of the Web - hard to identify and retrieve. Websites like Patient Opinion give patients and carers a new and powerful voice, allowing stories and experiences to be shared, feedback to be offered and new kinds of social reputation to be created.

To effectively distil, manage and process this social information in an automated way, we propose a novel process for the extraction of cognitive and affective information from patient opinions consisting of four main modules: a NLP module, which performs a first skim of the document, a Semantic Parser, whose aim is to extract concepts from the lemmatized text, the ConceptNet module, for the inference of the semantics associated with the given concepts, and AffectiveSpace, for the extraction of sentics (Fig. 2).

The NLP module interprets all the affective valence indicators usually contained in text such as special punctuation, complete upper-case words, onomatopoeic repetitions, exclamation words, negations, degree adverbs and emoticons, and eventually lemmatizes text.

The Semantic Parser then deconstructs text into concepts using a lexicon based on 'sentic n-grams' i.e. sequences of lexemes which represent multiple-word common sense and affective concepts extracted from the Open Mind corpus, WNA and other linguistic resources. The module also provides, for each retrieved concept, the relative frequency, valence and status, that is the concept's occurrence in the text, its positive or negative connotation and the degree of intensity with which the concept is expressed.

The ConceptNet module finds matches between the retrieved concepts and those previously calculated using CF-IOF and spectral association. In particular, CF-IOF weighting is exploited to find seed concepts for a set of a-priori categories, extracted from Patient Opinion, and spectral association is then used to expand this set with semantically related common sense concepts.
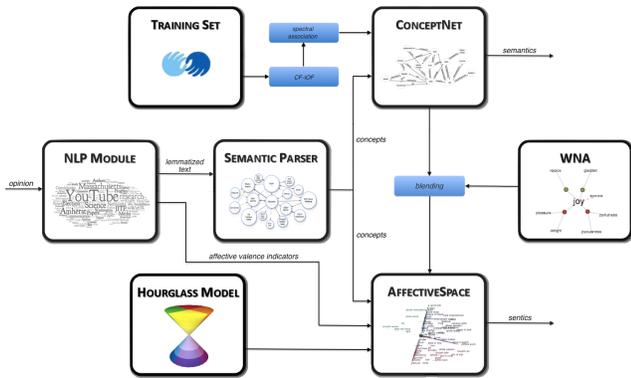
**Figure 2: Semantics and sentics extraction process**

The AffectiveSpace module projects the retrieved concepts into the vector space built by merging ConceptNet and WNA. The multi-dimensional space, clustered with respect to the Hourglass model, is then exploited to infer the affective valence of the retrieved concepts, in terms of Pleasantness, Attention, Sensitivity and Aptitude, according to the relative position they occupy in the space.

## 3. CROWD VALIDATION

Once natural language data are converted to a structured format, each topic expressed in each patient opinion and its related polarity can be aggregated and compared. We can then easily assimilate them with structured health-care information contained in a database or available through an API. We call this process 'crowd validation' [5], because the feedback comes from the masses, and we believe it will be the next frontier of health-care since patient opinions are crucial in understanding the effectiveness and efficiency of health services and how they can be improved.

Within this work, in particular, we use the opinion analysis process to marshal Patient Opinion's social information in a machine-accessible and machine-processable format and, hence, compare it with the official hospital ratings provided by NHS Choices [10] and each NHS trust. We use the inferred ratings to validate the information declared by the relevant health-care providers, crawled separately from each NHS trust website, and the official NHS ranks, extracted using NHS Choices API.

In order to provide patients with a common framework to compare, validate and select their health-care providers, we exploit the multi-faceted classification paradigm. Faceted classification allows the assignment of multiple categories to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, pre-determined and taxonomic order. This makes it possible to perform searches combining the textual approach with the navigational one. Faceted search, in fact, enables users to navigate a multi-dimensional information space by writing queries in a text box and progressively narrowing choices in each dimension.

In particular, to build the crowd validation interface, we use SIMILE Exhibit API [6], a set of Javascript files that enables the creation of rich interactive web-pages including maps, timelines and galleries, with very detailed client-side filtering.

Exhibit pages use the multi-faceted classification paradigm to display semantically structured data stored in a Semantic Web aware format e.g. RDF or JavaScript object notation (JSON). One of the most relevant aspects of Exhibit is that, once the page is loaded, the web-browser also loads the entire data set in a lightweight database and performs all the computations (sorting, filtering, etc.) locally on the client-side, providing high performances (Fig. 3).

The information contained in the Sesame triple-store, together with the information declared by each health-care provider (crawled separately from the relevant NHS trust website) and the official NHS ranks (extracted using NHS Choices API), is exported to a JSON file in order to feed the Exhibit interface and, hence, make the data available for browsing as an interconnected knowledge base.

NHS trusts are displayed in a dynamic gallery that can be ordered according to different parameters, either textual or numeric, concerning both general information about the health-care provider (e.g. location, services, facilities, transport and opening hours) and service quality of each NHS trust (e.g. communication, food, parking, service, staff and timeliness).

The service ratings, in particular, can be dynamically browsed and compared according to the different sources (NHS trust, NHS Choices or Patient Opinion) with relative links for accessing the original information. Thanks to faceted menus, it is possible to explore such information both by using the search box (to perform keyword-based queries) and by filtering the results through constraint addition or retraction on the facet properties.

## 4. EVALUATION

As a preliminary evaluation of the system, we tested crowd validation's capability to extract cognitive and affective information from patient opinions and performed some first usability tests. In particular, in order to calculate statistical classifications such as precision and recall of the semantics and sentics extraction process, we evaluated the system with a corpus of topic and mood tagged blogs from LiveJournal (LJ) [9], a virtual community of more than 23 million users who keep a blog, journal or diary. One of the interesting features of this website is that LJ bloggers are allowed to label their posts not only with a topic tag but also with a mood label, by choosing from more than 130 predefined moods or by creating custom mood themes.

Since the indication of the affective status is optional, the mood-tagged posts are likely to reflect the true mood of the authors and, hence, form a good test-set for crowd validation. As for the topic tags, in turn, we selected the LJ labels that match Patient Opinion topic-tags e.g. 'food', 'cleanliness' or 'communication', in order to collect natural language text that is likely to have the same semantics as the cognitive information usually associated with patient opinions. After retrieving and storing relevant data and metadata from 10,000 LJ posts, we extracted semantics and sentics through the opinion analysis process and compared the output with the relative topic and mood tags, in order to calculate precision, recall and F-measure rates.

On average, each post contained around 140 words, from which about 12 affective valence indicators and 60 concepts were extracted. From the retrieved concepts we inferred semantics and sentics associated with each of the selected posts and, hence, tagged them with topic and mood labels.

We then compared these labels with the corresponding topic and mood LJ tags, obtaining very good accuracy in terms of both semantics and sentics extraction. As for the detection of moods, for example, 'happy' and 'sad' posts were identified with a precision of 89% and 81% and recall rates of 76% and 68% respectively, that is with total F-measure values of 82% and 74%. As for the detection of topics, in turn, the classification of 'food' and 'communication' posts was performed with a precision of 75% and 69% and recall rates of 65% and 58% respectively. The total F-measure rates, hence, were considerably good (70% for 'food' posts and 63% for 'communication' posts)

We also performed some first usability tests in order to evaluate performances of the crowd validation interface and its user-friendliness. Search and retrieval tests on NHS trusts were performed on a group of 10 expert web-users. Users were asked to freely browse information about NHS services using the crowd validation IUI and to retrieve information about specific health-care providers, in order to judge both usability and accuracy of the interface. What emerged from these preliminary tests is that users really appreciate being able to dynamically and quickly set/remove constraints in order to display specific information about NHS trusts.

Patient Opinion semantic and sentic facets, in particular, were mostly used by participants for judging the quality of the different health-care providers. Users also really appreciated the possibility of consulting information from the original source (e.g. reading a particular patient opinion about a specific topic). In the near future, we plan to carry out broader usability tests and to check the validity of the discrepancy between official and unofficial ratings by manually assessing the more controversial values. Further results will be submitted elsewhere for publication.

## 5. CONCLUSION AND FUTURE WORK

Medicine is finally waking up to the use of social networking to listen to the 'wisdom of the patient'. Health-care of the future will be based on community, collaboration, self-caring, co-creation and co-production using technologies delivered via the Internet. This shift in emphasis to e-health does not replace traditional health care models but rather complements them and will ideally become the prevailing model. The Internet promises to be the 'silver bullet' to allow medicine to be affordable in the 21st century. Using the Internet as a platform, new models of proactive care complementing the current reactive paradigm are beginning to enter mainstream medicine.

From this perspective, on-line patient opinions will be central to understanding the effectiveness and efficiency of services and how they can be improved. To bridge the gap between this social information and the structured data supplied by health-care providers, we exploited both the semantics and sentics associated with patient opinions over the Web and, hence, provided the real end-users of the health system with a common framework to compare, validate and select their health-care providers.

We are currently working on making the system adaptive. We plan to assign to every piece of information stored in the crowd validation database a confidence score, which will be increased/decreased according to users' feedback on health-care ratings. We also plan to improve the IUI by adding new functionalities, such as the option to display NHS trusts according to their location caption on a map or



Figure 3: Multi-faceted interface

to compare two or more health-care providers together according to their overall ratings. We feel that Web 2.0 has a great potential that still needs to be exploited, especially in the field of e-health, and we believe that bridging the gap between structured and unstructured health-care data through semantics and sentics is the way to proceed towards the development of next-generation Health Web Science applications.

## 6. REFERENCES

[1] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro. Sentic Computing for Patient Centered Applications. In *Proceedings of IEEE ICSP10*, Beijing, 2010.

[2] E. Cambria, A. Hussain, C. Havasi, and C. Eckl. AffectiveSpace: Blending Common Sense and Affective Knowledge to Perform Emotive Reasoning. In *WOMSA09*, Seville, 2009.

[3] E. Cambria, A. Hussain, C. Havasi, and C. Eckl. Sentic Computing: Exploitation of Common Sense for the Development of Emotion-Sensitive Systems. volume 5967 of *Lecture Notes in Computer Science*, pages 148–156. Springer, Berlin Heidelberg, 2010.

[4] E. Cambria, A. Hussain, C. Havasi, and C. Eckl. SenticSpace: Visualizing Opinions and Sentiments in a Multi-Dimensional Vector Space. volume 6279 of *Lecture Notes in Computer Science*, pages 385–393. Springer, Berlin Heidelberg, 2010.

[5] E. Cambria, A. Hussain, C. Havasi, C. Eckl, and J. Munro. Towards Crowd Validation of the UK National Health Service. In *Proceedings of WebSci10*, Raleigh, NC, 2010.

[6] Exhibit. http://simile-widgets.org/exhibit, 2011.

[7] C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Recent Advances in Natural Language Processing*, Borovets, 2007.

[8] C. Havasi, R. Speer, and J. Holmgren. Automated Color Selection Using Semantic Knowledge. In *Common Sense Knowledge: Papers from the AAAI Fall Symposium*, Arlington, 2010.

[9] LiveJournal. http://livejournal.com, 2011.

[10] NHSChoices. http://www.nhs.uk, 2011.

[11] PatientOpinion. http://patientopinion.org.uk, 2011.

[12] Sesame. http://openrdf.org, 2009.

[13] C. Strapparava and A. Valitutti. WordNet-Affect: an Affective Extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086, 2004.