SEMAPRO 2017

The Eleventh International Conference on Advances in Semantic Processing

ISBN: 978-1-61208-600-2

November 12 - 16, 2017

Barcelona, Spain

SEMAPRO 2017 Editors

Wladyslaw Homenda, Warsaw University of Technology, Poland

Dumitru Roman, SINTEF/University of Oslo, Norway

SEMAPRO 2017

Forward

The Eleventh International Conference on Advances in Semantic Processing (SEMAPRO 2017), held between November 12 - 16, 2017, in Barcelona, Spain, continued a series of events related to the complexity of understanding and processing information.

Semantic processing considers contextual dependencies and adds to the individually acquired knowledge emergent properties and understanding. Hardware and software support and platforms were developed for semantically enhanced information retrieval and interpretation. Searching for video, voice and speech [VVS] raises additional problems to specialized engines with respect to text search. Contextual searching and special patterns-based techniques are current solutions.

With the progress on ontology, web services, semantic social media, semantic web, deep web search /deep semantic web/, semantic deep web, semantic networking and semantic reasoning, SEMAPRO 2017 constitutes the stage for the state-of-the-art on the most recent advances.

The event was very competitive in its selection process and very well perceived by the international scientific and industrial communities. As such, it has attracted excellent contributions and active participation from all over the world. We were very pleased to receive a large number of top quality contributions.

The conference had the following tracks:

- Basics on semantics
- Semantic applications/platforms/tools

We take here the opportunity to warmly thank all the members of the SEMAPRO 2017 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to SEMAPRO 2017. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also gratefully thank the members of the SEMAPRO 2017 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that SEMAPRO 2017 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the field of advanced semantic processing. We also hope that Barcelona, Spain, provided a pleasant environment during the conference and everyone saved some time to enjoy the unique charm of the city.

SEMAPRO 2017 Chairs

SEMAPRO Steering Committee

Fabio Grandi, University of Bologna, Italy
Tassilo Pellegrini, St. Pölten University of Applied Sciences, Austria
Sandra Lovrenčić, University of Zagreb, Croatia
Giuseppe Berio, Université de Bretagne Sud / IRISA, France
Takahiro Kawamura, Japan Science and Technology Agency (JST), Japan
Sule Yildirim Yayilgan, Gjøvik University College, Norway
Michele Melchiori, Università degli Studi di Brescia, Italy
Muhammad Javed, Cornell University, USA
Wladyslaw Homenda, Warsaw University of Technology, Poland

SEMAPRO Industry/Research Advisory Committee

Enrico Francesconi, ITTIG-CNR / Publications Office of the EU, Italy
Peera Pacharintanakul, TOT, Thailand
Mari Wigham, Wageningen Food & Biobased Research, The Netherlands
Raoul G. C. Schönhof, High Performance Computing Center Stuttgart (HLRS), Germany
Raghava Mutharaju, GE Global Research, Niskayuna, USA
Riccardo Albertoni, Istituto per la Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" - Consiglio Nazionale delle Ricerche (IMATI-CNR), Italy
Sofia Athenikos, Bank of America Merrill Lynch, USA
Shun Hattori, Muroran Institute of Technology, Japan

SEMAPRO 2017
Committee

SEMAPRO Steering Committee

Fabio Grandi, University of Bologna, Italy
Tassilo Pellegrini, St. Pölten University of Applied Sciences, Austria
Sandra Lovrenčić, University of Zagreb, Croatia
Giuseppe Berio, Université de Bretagne Sud / IRISA, France
Takahiro Kawamura, Japan Science and Technology Agency (JST), Japan
Sule Yildirim Yayilgan, Gjøvik University College, Norway
Michele Melchiori, Università degli Studi di Brescia, Italy
Muhammad Javed, Cornell University, USA
Wladyslaw Homenda, Warsaw University of Technology, Poland

SEMAPRO Industry/Research Advisory Committee

Enrico Francesconi, ITTIG-CNR / Publications Office of the EU, Italy
Peera Pacharintanakul, TOT, Thailand
Mari Wigham, Wageningen Food & Biobased Research, The Netherlands
Raoul G. C. Schönhof, High Performance Computing Center Stuttgart (HLRS), Germany
Raghava Mutharaju, GE Global Research, Niskayuna, USA
Riccardo Albertoni, Istituto per la Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" - Consiglio Nazionale delle Ricerche (IMATI-CNR), Italy
Sofia Athenikos, Bank of America Merrill Lynch, USA
Shun Hattori, Muroran Institute of Technology, Japan

SEMAPRO 2017 Technical Program Committee

Riccardo Albertoni, Istituto per la Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" -Consiglio Nazionale delle Ricerche (IMATI-CNR), Italy
Jose María Alvarez Rodríguez, Carlos III University of Madrid, Spain
Sofia Athenikos, Bank of America Merrill Lynch, USA
Agnese Augello, ICAR - Istituto di Calcolo e Reti ad alte prestazioni | Consiglio Nazionale delle Ricerche, Palermo, Italy
Isabel Azevedo, Instituto Superior de Engenharia do Porto (ISEP), Porto, Portugal
Fernanda Baiao, Federal University of the State of Rio de Janeiro, Brazil
Jarosław Bąk, Poznan University of Technology, Poland
Giuseppe Berio, Université de Bretagne Sud / IRISA, France
Jorge Bernardino, Polytechnic of Coimbra - ISEC, Portugal
Stefano Bortoli, HUAWEI TECHNOLOGIES Duesseldorf GmbH - German Research Center - Munich Office, Germany
Loris Bozzato, Fondazione Bruno Kessler, Trento, Italy

"Demokritos", Greece
Efstratios Kontopoulos, Information Technologies Institute (ITI) / Center for Research & Technology Hellas (CERTH), Greece
Jaroslav Kuchař, Czech Technical University in Prague, Czech Republic
Kyu-Chul Lee, Chungnam National University, Republic of Korea
Antonio Lieto, University of Turin and ICAR-CNR, Italy
Giuseppe Loseto, Polytechnic University of Bari, Italy
Sandra Lovrenčić, University of Zagreb, Croatia
Wencan Luo, University of Pittsburgh, USA
Miguel Felix Mata Rivera, UPIITA-IPN, Mexico
Brigitte Mathiak, GESIS - Leibniz institute for Social Sciences, Germany
John McCrae, Insight Centre for Data Analytics | National University of Ireland, Galway, Ireland
Imen Megdiche, Institut National Universitaire Champollion | IRIT, France
Muntazir Mehdi, Insight Centre for Data Analytics | National University of Ireland Galway, Ireland
Michele Melchiori, Università degli Studi di Brescia, Italy
Eleni Mikroyannidi, British Broadcasting Corporation (BBC), UK
Paramita Mirza, Max Planck Institute for Informatics, Germany
Panagiotis Mitzias, Information Technologies Institute (ITI) of the Centre for Research & Technology Hellas (CERTH), Greece
Fleur Mougin, University of Bordeaux, France
Diego Moussallem, University of Leipzig, Germany
Raghava Mutharaju, GE Global Research, Niskayuna, USA
Fatemeh Nargesian, University of Toronto, Canada
Lyndon Nixon, MODUL Technology GmbH, Austria
Peera Pacharintanakul, TOT, Thailand
Panagiotis Papadakos, FORTH-ICS (Foundation for Research and Technology - Hellas, Institute of Computer Science), Greece
Tassilo Pellegrini, St. Pölten University of Applied Sciences, Austria
Rivindu Perera, Auckland University of Technology, New Zealand
Livia Predoiu, University of Oxford, UK
Simon Razniewski, Free University of Bozen-Bolzano, Italy
Kate Revoredo, Federal University of the State of Rio de Janeiro - UNIRIO, Brazil
German Rigau Claramunt, University of the Basque Country, Spain
Juergen Rilling, Concordia University, Montreal, Canada
Tarmo Robal, Tallinn University of Technology, Estonia
Alejandro Rodríguez González, Universidad Politécnica de Madrid, Spain
Michele Ruta, Technical University of Bari, Italy
Minoru Sasaki, Ibaraki University, Japan
Raoul G. C. Schönhof, High Performance Computing Center Stuttgart (HLRS), Germany
Kinga Schumacher, German Research Center for Artificial Intelligence (DFKI GmbH), Germany
Wieland Schwinger, Johannes Kepler University Linz (JKU) | Inst. f. Telekooperation (TK), Linz, Austria

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Cognitive Map-Based Representation for Consumer Behaviour Modelling

Wladyslaw Homenda
Faculty of Economics and Informatics in Vilnius
University of Bialystok, Vilnius, Lithuania
Faculty of Mathematics and Information Science
Warsaw University of Technology, Warsaw, Poland
Email: homenda@mini.pw.edu.pl

Agnieszka Jastrzebska
Faculty of Mathematics and Information Science
Warsaw University of Technology, Warsaw, Poland
Systems Research Institute
Polish Academy of Sciences, Warsaw, Poland
Email: A.Jastrzebska@mini.pw.edu.pl

***Abstract*—Human behaviour modelling is a prominent example of an area in which information processing and understanding remains a challenge. Among crucial problems of this domain are heterogeneity, multiplicity and uncertainty of information. Behaviour modelling benefits from methods that enhance understanding of dependencies between phenomena and form a comprehensive model over a collection of elementary information granules. If we consider analytical application, quantitative predictive modelling becomes obsolete, because it is unable to represent wealth of information and its structuring. Hence, there is a need for semantic knowledge modelling. In light of the above, we present a Cognitive Map-based modelling framework capable to represent decision making processes. The model assumes that motivational stimuli determine decision making outcome. In case of human decision making, needs play the role of motivational stimuli. A decision is an outcome of processing of human needs. In order to reflect this using a Cognitive Map-based model, we assume that concepts making a map correspond to various needs. In the paper, we present a processing scenario that applies a Cognitive Map of needs and a current state of personal stimuli to produce a decision. We also apply the model to real-world data in an experiment of mobile phone activity monitoring.**

***Index Terms*—Cognitive Maps; Fuzzy Cognitive Maps; consumer behaviour modelling; decision making.**

I. Introduction

Human decision making elicits from an entangled collection of needs, and depends on the current state of mind and external conditions. Human decision making modelling is challenging both on the knowledge representation level and on processing level. First, human needs are not homogeneous, there are many of them, they are dependent one on another, and there is no objective and precise way to express their intensity and character. This is the case when applying a non-standard knowledge representation model that operates on abstract concepts. Particular implementation of a concept should be easy to interpret by a human being and it should relate to a corresponding need.

Even though needs are an internal collection of stimuli, they appear in some external context. In order to define a model reflecting the described assumptions, we require some variant of a semantic knowledge model with appropriate formalism for data processing.

We propose to apply Cognitive Maps to consumer decision making modelling. Cognitive Map is an example of a semantic knowledge model. It is a soft computing method composed of a collection of concepts and relationships between them. Processing with a Cognitive Map could be envisioned as follows: as the input, we provide data corresponding to a current state of phenomena. Map processes the input using concepts and dependencies between concepts and produces output data corresponding to the updated state of phenomena.

The key novelty introduced in this paper is the application of Cognitive Maps to consumer preferences and decision making modelling. So far, Cognitive Maps have been applied to: time series prediction on a linguistic level [1], pattern recognition [2] and system modelling and control [3].

The paper is structured as follows. Section II is a brief presentation of background knowledge on consumer decision making modelling. Section III introduces the new approach. Section IV addresses the method with the help of case studies. Section V concludes the paper.

II. Literature Review

In this section, we present a general overview of the most relevant streams of studies on consumer decision making modelling.

First, we need to mention the existence of a vast number of works on multi-criteria decision making. In this line of research, we come across methods for weighting and aggregating criteria relevant for a given decision. For instance, [4] addresses a multi-criteria decision making model utilising linguistic operators, while [5] addresses a method based on prioritised weighted aggregation with OWA (ordered weighted averaging) operator and t-norms. There is a wide range of papers in this domain employing various operators for aggregation and various models for knowledge representation. Among studies similar ours, we find methods that, apart from the task of criteria aggregation, take into account interactions between criteria. This, typically, is realised by assigning weights not only for individual criteria, but also to all combinations of criteria. Aggregation is performed for such extended formalism. An example of this approach based on Choquet integral playing the role of an aggregation function is presented in [6].

Another relevant group of studies revolves around preference modelling. Preference modelling is focused on comparing available alternatives. Let us assume we have objects x and

 ISBN: 978-1-61208-600-2

y from set X. We denote that x is at least as good as y as: $x \succsim y \Leftrightarrow f(x) \geqslant f(y)$, where $f : X \to \mathbb{R}$ is a valuation function that measures how good a given alternative is. In addition, in many applications a strict preference relation, denoted as $\succ$, is used. Apart from predominantly theoretical studies on preference relations, like [7], the literature offers impressive machine learning methods for preference learning. Their aim is to use training data to form a model that automatically performs alternatives' ranking. Here, we find SVM$^{\text{rank}}$ which solves an optimisation problem aiming at alternatives ranking [8], the ListNet [9] which is an algorithm based on neural network and gradient descent.

A Cognitive Map is a soft computing method forming a semantic knowledge model over a set of concepts. A map is a weighted digraph: it consists of concepts (vertices) and directed weighted edges linking the concepts. The philosophy of modelling with Cognitive Maps is very simple: concepts correspond to phenomena, linkages to relationships between phenomena. Weights inform about character and strength of connection. Weights are collected in a weight matrix. Such minimal formalisation encouraged extensive research on Cognitive Maps and led to the development of related sub-families, i.e., Fuzzy Cognitive Maps, Granular Cognitive Maps, etc. Versatility of the original formalism of Cognitive Maps proved to be so desirable and successful, that it has remained unchanged in all named sub-kinds of Cognitive Maps. What differentiates them is the assumed information representation model: crisp in Cognitive Maps, fuzzy in Fuzzy Cognitive Maps, granular in Granular Cognitive Maps, etc.

Intensive research on Cognitive Maps in application to knowledge processing started with a ground-breaking paper by B. Kosko [10]. The referenced paper presents a generalisation of Cognitive Maps to Fuzzy Cognitive Maps. More importantly, it presents a simple method for weight matrix learning. Nowadays, major studies on Cognitive Maps revolve around Fuzzy Cognitive Maps trained using a bio-inspired metaheuristic optimisation algorithm of choice. As mentioned, experimental fields, where Fuzzy Cognitive Maps have been successfully utilised include systems modelling and control, time series analysis, and pattern recognition.

In this paper, we present an application of Cognitive Maps to human decision making modelling, which is a novel and original contribution.

III. Preliminaries

A. Consumer Representation

The primary task is to define a model for consumer representation that will be the backbone of decision making modelling framework. Inspired by the research of K. Lewin on psychophysical field [11], we assume that a vector of all needs represents each consumer. Such infinite vector is a subject for further modelling. The vector in its most general form is given as:

$$\mathbf{x} = [x_1, x_2, \ldots]^T \tag{1}$$

$\mathbf{x}$ represents a given consumer, $x_i, i = 1, 2, \ldots, +\infty$ stands for his i-th need. It is worth to explain that a vector of needs $\mathbf{x}$ is of finite length in any practical application. Infinity in its description underlines that, despite its finiteness, we do not put restriction on its length, see below for details.

Each need in the vector is evaluated using a selected information representation scheme. In order to choose a particular model, one shall consider properties that are the most desired in a given application area. In our experiments, we considered crisp information and fuzzy sets [12], but there are other options available, for instance intuitionistic fuzzy sets.

In theory, the needs vector is infinite, because the set of needs is infinite. One aspect of human development is that it is accompanied by recognition of new needs. This phenomenon is explained by Maslow, [13], who states that after satisfying needs of a prime urgency, humans naturally start recognising more refined desires. Moreover, identification of new needs could be triggered by external stimuli, especially skilful marketing communication. The framework presented in this paper operates on a finite set of needs. This limitation is necessary, not only for computational reasons, but also because it is the only sensible convention to focus on needs relevant to a problem of interest. In practice, relevant needs must be identified before empirical data is collected. Let us give a few examples of sets of needs relevant in different decision making problems:

- needs relevant when we consider a particular car purchase: number of seats, capacity for carrying goods, ease of access to repair shops, level of extravagance associated with the car, etc.
- needs relevant when shopping for cleaning supplies: wood cleaner, glass cleaner, disinfectant, etc.
- needs related to an evening outing: concert, restaurant, theatre, visiting friends, etc.

Most importantly, the model allows describing causality. A particular needs vector describes preferences at a given at a given moment in time. Building consecutive vectors, with different needs evaluations will allow to represent time flow and illustrate change of needs.

The strength of motivational stimuli is expressed through needs evaluation. The method of evaluation depends on selected information representation model. If we employ classical set theory, a need either exists or it does not. Hence, evaluation of a need relies on selecting a number from the set $\{0, 1\}$. In contrast, with fuzzy set theory, needs are evaluated as real numbers from the interval $[0, 1]$. We can also perform needs evaluation based on linguistic variables or some other method we find suitable.

The model is capable to describe and discover dependencies at various levels of generality. This derives from the fact that we can easily group specific needs into more general clusters or the other way around: based on a general group of needs, we can transition into a fine-grained analysis, depending on the availability of data. For example, a general type of need

 ISBN: 978-1-61208-600-2

for existence includes all needs for food. The need for food contains a need for carbohydrates. The need for carbohydrates may be satisfied by consuming rice, bread, pasta, cookies, and so on. The structure is recursively nested.

Additionally, this relatively simple form of consumer representation allows sophisticated analysis, if we consider an abstract space of consumers described by their needs. In such a space, we may define classes of consumers and detect similarity between classes of consumers.

B. Processing with Cognitive Maps and Fuzzy Cognitive Maps

A Cognitive Map is a graph-based knowledge representation model. Let us denote the vertices in the map as $A_1, A_2, \ldots, A_c$, where c represents the number of vertices. Vertices correspond to abstract concepts. The strength of relationships between the vertices is denoted using weights: w_{11}, $w_{12}, \ldots w_{cc}$, where w_{ij} is a weight going from the concept A_i to the concept A_j. Weights are gathered in a $c \times c$ weight matrix denoted as $\mathbf{W}$. Particular values assumed by w_{ij} depend on an assumed information representation system (fuzzy, crisp, etc.).

In (regular) Cognitive Maps, weights assume values from the set $\{-1, 0, 1\}$. -1 indicates that an increase in a source node is correlated with a decrease in a destination node. 0 denotes lack of relationship. 1 informs that an increase in a source node causes increase in a destination node.

The constricted set of values allowed in regular Cognitive Maps limits their flexibility. The formulation of Fuzzy Cognitive Maps appeared as a remedy for this issue. In Fuzzy Cognitive Maps, relationships are expressed using real numbers from the $[-1, 1]$ interval [10].

Processing with any Cognitive Map is realised in the following way: as the input to the map we pass c activations gartered in a c-dimensional vector, where one activation is for one node in the map. Let us denote the vector of activations as $\mathbf{x} = [x_1, \ldots, x_c]^T$. Activations are the input data and they correspond to the state of nodes (the state of excitement of needs) at the current moment in time tm. The map processes the input using weights $\mathbf{W}$ and, as a result, we obtain the output. Let us denote the output as $\mathbf{y} = [y_1, \ldots, y_c]^T$. The output is interpreted as the state of nodes in the next moment in time $(tm + 1)$. In other words, the Cognitive Map models changes of a system of concepts in time. The idea behind processing with a Cognitive Map composed of three nodes is illustrated in Figure 1. Computations are formally represented as follows:

$$\mathbf{y} = \mathbf{W} \star \mathbf{x} \tag{2}$$

where $\star$ is an operation on a matrix and a vector.

Typically, but not always, in Cognitive Maps, input and output vectors are evaluated using values from the set $\{0, 1\}$, while Fuzzy Cognitive Maps use real numbers from the interval $[0, 1]$. Particular examples of implementations of $\star$ operation are given in Section IV.

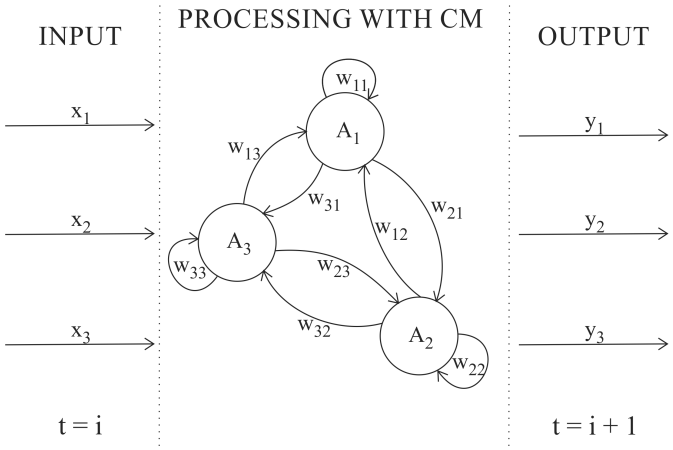

Fig. 1. Processing with a Cognitive Map (CM) with three nodes.

Map outputs represent state of nodes in the next moment in time. We wish that the predicted states are as close as possible to the actual states observable in future. In other words, maps allow predicting. The prediction quality is expressed by calculating similarity between map output $\mathbf{y}$ and desired target values, denoted as $\mathbf{t}$. This could be expressed as:

$$\mathbf{y} \sim \mathbf{t} \tag{3}$$

$\mathbf{t} = [t_1, t_2, \ldots, t_c]^T$ is the desired, ideal target.

The modelling outcome depends on the weight matrix. There are three strategies that could be executed in order to obtain a weight matrix:

- involve human experts to manually propose a weight matrix,
- run supervised learning procedure to extract a weight matrix automatically from data,
- hybrid learning: use expert knowledge to define fine-grained conditions to facilitate a better performance of a learning algorithm.

The first approach is the most inconvenient, because in order to build a map, we need human involvement and the procedure is manual. The substantial advantage of this approach is that when a Cognitive Map is constructed by experts, its formalism is well understood. Humans provide a structured representation of phenomena in a way that is easy to interpret and conveys a meaningful description.

Automated weight matrix learning has the big advantage that it does not require the considerable effort needed when we involve human experts. In this case, we use historical data to obtain a weight matrix. In practice, there are a few obstacles in this approach. First, not all applications have enough historical data. Second, trends in development of algorithms have brought forward nature-inspired optimisation heuristics, like Genetic Algorithm, Particle Swarm Optimisation, and so on. They are applied in areas when optimisation is difficult, as this one, and their motto is (so to speak) “close enough is good enough”. More formally speaking, they do not guarantee convergence to the optimum. In many domains

 ISBN: 978-1-61208-600-2

of applied Cognitive Maps, for instance in classification, this property does not affect the modelling outcome as weights of connections between map nodes do not matter as much as the outcome, which is, an assigned class label. In contrast, if we want to obtain a meaningful knowledge about relationships between the nodes, the "close enough is good enough" motto is not what we shall be content with. It might happen that two, substantially different maps (meaning: maps with drastically different evaluations of weights) could provide similar numerical modelling accuracy.

C. Cognitive Map-based Consumer Needs Representation

In our approach to human decision making modelling, we assume that needs determine actions. We apply the field theory of K. Lewin, who is recognised as the founder of social psychology, [14]. The field theory can be transparently transformed into a mathematical model. This is a rare and valuable property, because many psychological theories are rather descriptive than quantitative. The field theory says that at a given moment in time we exist within or, to put it in another words, we own a certain abstract psychophysical field comprising of all needs there are. The strengths of needs in a psychophysical field change over time. The reason for change could be internal or external. There is also an assumption that any part of a psychophysical field depends on its every other part. Human behaviour can be explained by analysing forces acting in the field.

Our model captures the moment when needs get re-evaluated in response to new input conditions. Hence, we assume that:

- a Cognitive Map represents relationships between human needs,
- input data (activations) of the Cognitive Map correspond to the current strength of motivational stimuli,
- the Cognitive Map's output represents strength of needs in response to the input, analysing the output allows to make a decision.

The premise behind introducing Cognitive Maps to represent consumer needs was dictated by review of the literature on needs taxonomies. We see an insufficient focus on methods that take into account relationships between the needs.

In future, we plan to impose certain conditions on the needs model. In this case, the most advanced, and at the same time interpretable, form of an arbitrarily defined set of conditions is an ontology of needs.

IV. Case Study

In this section, let us introduce a few examples of application of Cognitive Maps to consumer decision making and preference modelling.

Two case studies are arbitrarily defined. They concern decision making modelling using a regular Cognitive Map and a Fuzzy Cognitive Map. The discussion is limited to relatively small maps representing five needs:

- listening to a radio,
- watching a TV,
- reading a book,
- playing with a dog,
- taking a walk.

The third case study presents another application of the model - to consumer space modelling. We apply a Fuzzy Cognitive Map to process a dataset describing mobile network activity.

A. Cognitive Map for Decision-Making Modelling

The first example covers the most basic version of the model. Is is based on a (regular) Cognitive Map, in which relationships between the nodes are expressed as values from the set $\{-1, 0, 1\}$. Input activations assume values from the set $\{0, 1\}$. We analyse a Cognitive Map for one consumer. Relationships between needs were defined arbitrarily and are displayed in Figure 2.

Activations describe current excitement levels of the needs under consideration. Let us assume, that the consumer reports the following activations: $\mathbf{x} = [0; 0; 1; 1; 0]^T$. The activations order corresponds to the list mentioned above. We interpret it as follows: lack of need for listening to a radio, watching a TV and taking a walk; existing need for reading a book and playing with a dog.

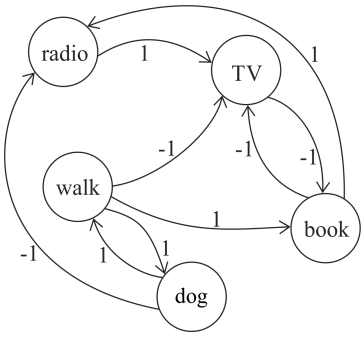

Fig. 2. Arbitrarily defined Cognitive Map for the case study consumer.

We propose to apply the following, very simple, scheme to implement the $\star$ operation, cf. (2). A single output y_i is calculated as follows:

$$y_i = \sum_{j=1}^{c} w_{ij} \cdot x_j \tag{4}$$

c denotes the number of nodes in the map. In this case, $c = 5$. For simplicity, we do not introduce any scaling to the outcome of the sum of products. Therefore, y_i assumes a value from the set $\{-c, -(c-1), \ldots, 0, \ldots, c-1, c\}$. The Maximum value in the output vector indicates a decision. It is very easy to imagine more sophisticated aggregation schemes.

Let us present the decision making modelling based on the stated assumptions:

$$\begin{matrix} radio \\ TV \\ book \\ dog \\ walk \end{matrix} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 1 & 0 \end{bmatrix} \star \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

In response to the presented activations, the need to go for a walk will be perceived as the strongest.

B. Fuzzy Cognitive Map for Decision Making Modelling

In the second case study, we present a Fuzzy Cognitive Map constructed for one consumer concerning the same set of five needs as in the previous example. The map was defined arbitrarily and is displayed in Figure 3.

Let us recall that in Fuzzy Cognitive Maps weights are evaluated as numbers from the $[-1, 1]$ interval, while activations and outputs are numbers from the $[0, 1]$ interval. The implementation of the $\star$ operator from (2) on the level of a single node is realised as follows:

$$y_i = f\big(\sum_{j=1}^{c} w_{ij} \cdot x_j\big) \tag{5}$$

for $i = 1, 2, \ldots, c$. f is a squashing function, which draws the product to the $[0, 1]$ interval. The sigmoid function with a steepness parameter $\tau > 0$ is most commonly used.

$$f(u) = \frac{1}{1 + e^{-\tau u}} \tag{6}$$

We assumed $\tau = 5$ based on literature [15].

Fuzzy Cognitive Maps allow greater flexibility in expressing preferences and relationships between needs.

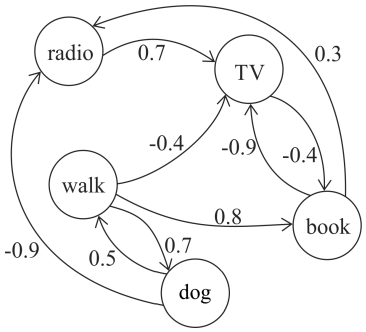

Fig. 3. Arbitrarily proposed Fuzzy Cognitive Map for the case study.

Let us assume, that current activations for the consumer, for whom we defined the Fuzzy Cognitive Map, are as follows: $\mathbf{x} = [0.1; 0; 0.4; 0.1; 0.6]^T$. The activations vector $\mathbf{x}$ informs us that out of five considered options (radio, TV, book, playing with a dog, and walking) the consumer has the strongest urge to go for a walk, watching TV is a non-existing stimulus, playing with a dog and listening to a radio are a couple of very weakly recognised needs, reading a book is a weakly moderate need. Let us present the computations in this case:

$$\begin{matrix} radio \\ TV \\ book \\ dog \\ walk \end{matrix} \begin{bmatrix} 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & -0.4 & 0 & 0 \\ 0.3 & -0.9 & 0 & 0 & 0 \\ -0.9 & 0 & 0 & 0 & 0.5 \\ 0 & -0.4 & 0.8 & 0.7 & 0 \end{bmatrix} \star \begin{bmatrix} 0.1 \\ 0 \\ 0.4 \\ 0.1 \\ 0.6 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.3 \\ 0.5 \\ 0.7 \\ 0.9 \end{bmatrix}$$

The modelling outcome, $\mathbf{y} = [0.5; 0.3; 0.5; 0.7; 0.9]^T$, informs us that taking a walk is the most likely choice for this person. The need for walking is the strongest at the computed moment in time. The second strongest need in the output vector $\mathbf{y}$ is the need for playing with a dog computed as 0.7. A strong positive connection, that joins walking with playing with a dog, caused the strength of this need to increase from 0.1 to 0.7.

C. The Mobile Activity Dataset

Preferences analysis concerns not only heterogeneous, but also homogeneous markets. It is a common practice to perform segmentation of heterogeneous markets into homogeneous groups. Further analysis of consumer preferences within one, homogeneous segment is an important assignment. An example of a homogeneous good is mobile network service. Importantly, statistics collected by mobile network service providers contain significant valuable information about cities and their citizens. It has been proposed to use subscriber's location statistics to traffic monitoring in public transportation services, [16]. Public transportation services, alike mobile services, are homogeneous goods, especially when we consider large cities.

We propose to apply a Fuzzy Cognitive Map to model and predict potential demand for public transportation services based on mobile subscriber's location statistics. We process data collected as a part of the VaVeL project (VaVeL: Variety, Veracity, VaLue [17]), concerning traffic in mobile networks recorded by Base Transceiver Stations in Warsaw. The data was collected hourly, starting from 0:00 on 9$^{\text{th}}$ January 2017, ending at 23:00 on 20$^{\text{th}}$ January 2017. We scaled it to the interval $[0.1, 0.9]$ so that the training procedure for a Fuzzy Cognitive Map can be conducted without numerical problems.

The course of experiment was as follows. First, we arbitrarily selected data from four neighbouring zones, out of the total of 895 geographical zones in the city. For each zone, we have information (a time series) concerning the number of mobile network subscriber's registered at each hour. We assume the sliding window model with three time points within one window. Based on these assumptions, we form a Fuzzy Cognitive Map based on 12 concepts, four zones multiplied by three moments in time (length of sliding window) gives us 12. The first three concepts correspond to the first zone, the next three concepts correspond to the second zone, etc. We formed activations and targets to represent changes of concepts states in time (in other words, changes of traffic). Predictions are for one step ahead. The complete methodology of this approach was discussed in [18]. There are

 ISBN: 978-1-61208-600-2

two alternative methods of how to use a sliding window Fuzzy Cognitive Map model. The first is that we average the model responses and produce a time series prediction in the form of a sequence of numbers. In this case a particular prediction is one point. The second method, more suitable for this study, is when we obtain a prediction in the form of a sequence of intervals. In this case a particular prediction is in the form of a lower and an upper interval limit, between which we assume that the state of phenomena is acceptable.

We run Particle Swarm Optimization to determine the best weights by minimization of Mean Squared Error between the Fuzzy Cognitive Map responses and target values:

$$MSE = \frac{1}{N \cdot c} \sum_{i=1}^{N} \sum_{j=1}^{c} (y_{ji} - t_{ji})^2 \quad (7)$$

N is the number of samples, c is the number of nodes, y are map responses, t are targets. We used 160 pairs of activation and target for map training and the remaining 80 pairs were left for model testing. We can technically perceive the data used for map training as a time series of four variables.

In addition, in the application of traffic monitoring, it is worth to consider softening decision rules. In particular, we can extend the interval of acceptable values by some reasonable value. An example of a measure that can be used to evaluate the degree to which we expand margins is standard deviation.

In order to verify if we trained a correct model we plot predicted interval limits and actual values of time series for the four studied zones. This is presented in Figure 4 for train and test sets. We soften the decision margin by adding half of standard deviation for each variable. The four variables have the following standard deviations: 0.1742, 0.2526, 0.1111, 0.0523. The predicted lower (green) and upper (red) limits for the four zones are illustrated in Figure 4. Black lines present true levels of traffic. The MSE for the four studied zones is present in each plot. We can conclude that the model was properly trained.

After the Fuzzy Cognitive Map was trained and we verified that it correctly describes the given data, we use it as a decision making aid. The application of the Fuzzy Cognitive Map can be informally described as follows. A person is planning to travel through the four zones. He collected information about the traffic in the last three hours in these zones and passes it as the input to the map. The map calculates a pessimistic and an optimistic prediction for the next moment in time. The person can decide whether to go or not to go.

Since the above application is trivial, we additionally tested the model in two less standard scenarios:

- How does the Map react if the input is consistently distorted? Consistently distorted input corresponds to a continuous sequence of recorded and known anomalous states of phenomena.

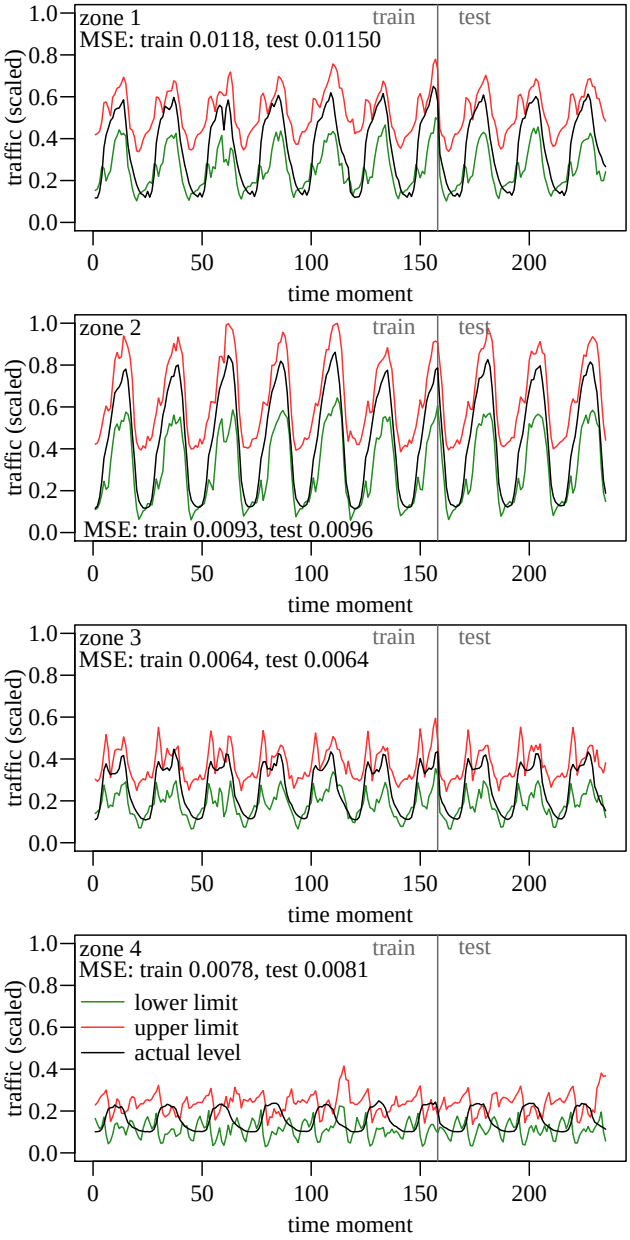

Fig. 4. Time series predictions for train and test sets in investigated zones.

- Is the Map sensitive to singular anomalies? Singular anomalies appear as single events, they are not preceded by anomalous events.

In Table I we present average coverage of train and test set observations by predicted intervals. The notion of coverage is quite straightforward, namely it is the number of true observations falling into predicted intervals. Intervals are spanned by the factor of a fraction of standard deviation. Without surprise, the more we expand the interval, the greater the coverage gets.

An analogous experiment was conducted for test data distorted with systematic noise. For this experiment, we added or subtracted a constant value to or from all activation vectors. We run activations with the Fuzzy Cognitive Map and we computed the average coverage of map outputs. The first test

 ISBN: 978-1-61208-600-2

TABLE I. AVERAGE COVERAGE OF TRAIN/TEST TIME SERIES DATA POINTS FOR PREDICTIONS WITHOUT AND WITH A SPANNING FACTOR

	interval spanning factor					
	zero		+0.2sd		+0.5sd	
zone	train	test	train	test	train	test
z1	0.3987	0.3506	0.5633	0.3506	0.7278	0.7403
z2	0.2848	0.2468	0.6392	0.2857	0.9304	0.8831
z3	0.4114	0.4026	0.5886	0.4805	0.7532	0.7532
z4	0.2215	0.2208	0.2721	0.2468	0.3734	0.3766

TABLE II. COVERAGE FOR TEST DATA MODIFIED WITH SYSTEMATIC DISTORTIONS

	interval spanning factor		
zone	zero	+0.2sd	+0.5sd
z1	0.1169	0.2468	0.3117
z2	0.2208	0.2987	0.3766
z3	0.1429	0.2078	0.3377
z4	0.1818	0.1948	0.2468

was performed for the original model, with the lower and the upper interval limits produced by the map. The next two tests were performed for intervals expanded by constant value equal to 0.2 of standard deviation and by 0.5 of standard deviation. The results are displayed in Table II. Without surprise, coverage is smaller, in comparison with results in Table I. Consistently, as we expand intervals, coverage grows.

Further experiments were for data that simulated unexpected anomalies. Anomalies were added randomly to target data. We can visually verify susceptibility of the trained Map to random distortions. The results (for a couple of zones) are illustrated in Figure 5. Random anomalies were generated by adding positive numbers drawn randomly from normal distribution with mean equal to 0 and standard deviation equal to 0.3. We added anomalies to 30 randomly selected values from the test set. Figure 5 concerns predictions expanded by 0.5 of standard deviation.

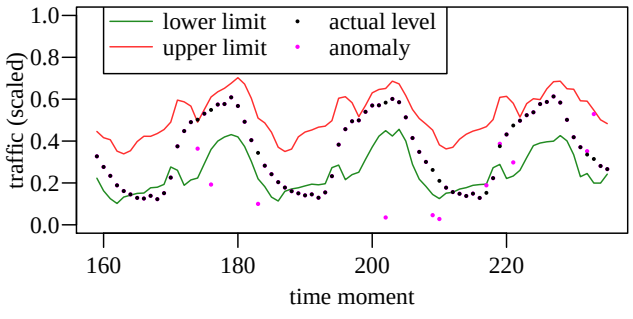

Fig. 5. Predictions for zone 1 test set with added anomalies.

V. CONCLUSION

In the paper, we have introduced a consumer decision making modelling approach based on Cognitive Maps and a vector-based representation of consumer needs. The method is flexible. Not only does it allow to model decision making processes, but also structures in consumer and needs spaces. The approach is on one hand easy to interpret, as a map provides a semantic knowledge representation, and, on the other hand, powerful, as the weight matrix can be automatically trained from historical data.

The intention of this paper was to introduce the idea of a versatile model for consumer preferences representation and modelling. The model requires further development and experimenting. We would like to emphasise that our study on homogeneous preferences analysis and the mobile activity dataset are in a very early stage. In future, we plan to analyse this dataset to a greater extent.

ACKNOWLEDGMENT

This research has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688380 *VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors*.

REFERENCES

[1] W. Stach, L. Kurgan, and W. Pedrycz, "Numerical and linguistic prediction of time series with the use of fuzzy cognitive maps," vol. 16, pp. 61 – 72, 2008.
[2] G. A. Papakostas, Y. S. Boutalis, D. E. Koulouriotis, and B. G. Mertzios, "Fuzzy cognitive maps for pattern recognition applications," vol. 22, no. 8, pp. 1461 – 1486, 2008.
[3] P. P. Groumpos and C. D. Stylios, "Modelling supervisory control systems using fuzzy cognitive maps," *Chaos, Solitons & Fractals*, vol. 11, no. 1, pp. 329 – 336, 2000.
[4] M. Aggarwal, "Adaptive linguistic weighted aggregation operators for multi-criteria decision making," *Applied Soft Computing*, vol. 58, pp. 690 – 699, 2017.
[5] H.-B. Yan, V.-N. Huynh, Y. Nakamori, and T. Murai, "On prioritized weighted aggregation in multi-criteria decision making," *Expert Systems with Applications*, vol. 38, no. 1, pp. 812 – 823, 2011.
[6] V. Cancer, "Considering interactions among multiple criteria for the server selection," vol. 34, no. 1, pp. 55 – 65, 2010.
[7] M. Roubens and P. Vincke, *Preference Modeling*. LNEMS 250, Springer Verlag, Berlin, 1985.
[8] T. Joachims, "Training linear svms in linear time," in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.
[9] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proceedings of the 24th International Conference on Machine Learning*, 2007.
[10] B. Kosko, "Fuzzy cognitive maps," vol. 24, pp. 65 – 75, 1986.
[11] K. Lewin, *Field Theory in Social Science*, D. Cartwright, Ed. Harper & Brothers, 1951.
[12] L. A. Zadeh, "Fuzzy sets," vol. 8, no. 3, pp. 338 – 353, 1965.
[13] A. Maslow, "A theory of human motivation," vol. 50, no. 4, pp. 370 – 396, 1943.
[14] S. J. Haggbloom, J. E. Warnick, V. K. Jones, G. L. Yarbrough, T. M. Russell, C. M. Borecky, R. Mcgahhey, J. L. I. Powell, J. Beavers, and E. Monte, "The 100 most eminent psychologists of the 20th century," vol. 6, no. 2, pp. 139 – 152, 2002.
[15] W. Stach, L. Kurgan, and W. Pedrycz, "A survey of fuzzy cognitive map learning methods," pp. 71 – 84, 2005.
[16] M. Luckner, A. Roslan, I. Krzeminska, J. Legierski, and R. Kunicki, "Clustering of mobile subscriber's location statistics for travel demand zones diversity," in *LNCS*, K. Saeed, W. Homenda, and R. Chaki, Eds., vol. 10244, 2017, paper 315 – 326.
[17] Vavel: Variety, veracity, value: Handling the multiplicity of urban sensors. [Online]. Available: http://www.vavel-project.eu/
[18] A. Jastrzebska and W. Homenda, "Modeling time series with fuzzy cognitive maps," in *Proc. of the IEEE World Congress on Computational Intelligence*, 2014, pp. 2572 – 2579.

 ISBN: 978-1-61208-600-2

Deduction System for Decision Logic based on Partial Semantics

Yotaro Nakayama* ‡, Seiki Akama† and Tetsuya Murai‡
*Nihon Unisys, Ltd., 1-1-1, Toyosu, Koto-ku, Tokyo, 135-8560, Japan
Email: yotaro.nakayama@unisys.co.jp
†C-Republic, Inc., 1-20-1 Higashi-Yurigaoka, Asao-ku, Kawasaki-shi, 215-0012, Japan
Email: akama@jcom.home.ne.jp
‡Chitose Institute of Science and Technology,
758-65 Bibi, Chitose, Hokkaido, 066-865, Japan
Email: t-murai@photon.chitose.ac.jp

***Abstract*—Rough set theory has been extensively used both as a mathematical foundation of granularity and vagueness in information systems and in a large number of applications. However, the decision logic for rough sets is based on classical bivalent logic; therefore, it would be desirable to develop decision logic for uncertain or ambiguous objects. In this study, a deduction system based on partial semantics is proposed for decision logic. Three-valued logics based on Gentzen sequent calculi are adopted. A deductive system based on three-valued framework is intuitively adequate for the structure of positive, negative, and boundary regions of rough sets, and has already been studied. In this study, consequence relations based on partial semantics for decision logic are defined, and systemization by Gentzen ' s sequent calculi is attempted. Three-valued logics of different structures are investigated as the deductive system of decision logic. The interpretation of decision logic is extended using partial semantics, and extended decision logic based on three-valued logics is proposed.**

Keywords–rough set; decision logic; consequence relation; three-valued logic; sequent calculi.

I. INTRODUCTION

Pawlak introduced the theory of rough sets for handling rough (coarse) information [1]. Rough set theory is now used as a mathematical foundation of granularity and vagueness in information systems and is applied to a variety of problems. In applying rough set theory, decision logic was proposed for interpreting information extracted from data tables. However, decision logic adopts the classical two-valued logic semantics. It is known that classical logic is not adequate for reasoning with indefinite and inconsistent information. Moreover, the paradoxes of material implication of classical logic are counterintuitive.

Rough set theory can handle the concept of approximation by the indiscernibility relation, which is a central concept in rough set theory. It is an equivalence relation, where all identical objects of sets are considered elementary. Rough set theory is concerned with the lower and the upper approximation of object sets. This approximation divides sets into three regions, namely, the positive, negative, and boundary regions. Thus, Pawlak rough sets have often been studied in a three-valued logic framework because the third value is thought to correspond to the boundary region of rough sets [2][3].

In this study, non-deterministic features are considered the characteristic of partial semantics. The formalization of three-valued logic is carried out using a consequence relation based on partial semantics. The basic logic for decision logic is assumed to be many-valued, in particular, three-valued and some of its alternatives [4]. If such three-valued logics are used as a basic deduction system for decision logic, it can be enhanced to a more useful method for data analysis and information processing. The decision logic of rough set theory will be axiomatized using Gentzen sequent calculi and three-valued semantic relation as basic theory. To introduce three-valued logic to decision logic, consequence relations based on partial interpretation are investigated, and sequent calculi of three-valued logic based on them are constructed. Subsequently, three-valued logics with different structure are considered for the deduction system of decision logic.

The deductive system of decision logic has been studied from the granule computing perspective, and in [5], an extension of decision logic was proposed for handling uncertain data tables by fuzzy and probabilistic methods. In [6], a natural deduction system based on classical logic was proposed for decision logic in granule computing. In [2], Gentzen-type three-valued sequent calculi were proposed for rough set theory based on non-deterministic matrices for semantic interpretation.

The paper is organized as follows. In Section Ⅱ, an overview of rough sets and decision logic is presented. In Section Ⅲ, the relationship between decision logic and three-valued semantics based on partial semantics is discussed. In Section Ⅳ, an axiomatization using Gentzen sequent calculus is presented, according to a consequence relation based on the previously discussed partial semantics. In Section V, an extension of decision logic is discussed, based on three-valued sequent calculus as partial logic. Finally, in Section Ⅵ, a summary of the study and possible directions for future work are provided.

II. OVERVIEW OF ROUGH SETS AND DECISION LOGIC

Rough set theory, proposed by Pawlak [1], provides a theoretical basis of sets based on approximation concepts. A rough set can be seen as an approximation of a set. It is denoted by a pair of sets, called the lower and upper approximation of the set. Rough sets are used for imprecise data handling. For the upper and lower approximations, any subset X of U can be in any of three states, according to the membership relation of objects in U. If the positive and negative regions on a rough set are considered to correspond to the truth value of a logical form, then the boundary region corresponds to ambiguity in deciding truth or falsity. Thus, it is natural to adopt a three-valued logic.

 ISBN: 978-1-61208-600-2

Rough set theory is outlined below. Let U be a non-empty finite set, called a universe of objects. If R is an equivalence relation on U, then U/R denotes the family of all equivalence classes of R, and the pair (U, R) is called a Pawlak approximation space. A knowledge base K is defined as follows:

Definition 1. A knowledge base K is a pair $K = (U, R)$, where U is a universe of objects and **R** is a set of equivalence relations on objects in U.

Definition 2. Let $R \in \mathbf{R}$ be an equivalence relation of the knowledge base $K = (U, R)$, and X any subset of U. Then, the lower and upper approximations of X for R are defined as follows:

$$\underline{R}X = \bigcup\{Y \in U/R \mid Y \subseteq X\} = \{x \in U \mid [x]_{\mathrm{R}} \subseteq X\}$$
$$\overline{R}X = \bigcup\{Y \in U/R \mid Y \cap X \neq 0\} = \{x \in U \mid [x]_{\mathrm{R}} \cap X \neq \emptyset\ \}$$

Definition 3. If K = (U,R), R ∈ R, and X ⊆ U, then the R-positive, R-negative, and R-boundary regions of X with respect to R are defined respectively as follows:

$$POS_R(X) = \underline{R}X$$
$$NEG_R(X) = U - \overline{R}X$$
$$BN_R(X) = \overline{R}X - \underline{R}X$$

Let C and D be subsets of an attribute A, denoted as $C, D \subseteq A$. Moreover, it is assumed that C is a conditional attribute and D a decision attribute. Then, the decision table T is denoted by $T = (U, A, C, D)$.

The function $s_x : A \to V$ (for simplicity, the subscript x will be omitted) is defined where $\forall x \in U$, and $\forall a \in C \cup D$.

Language of Decision Logic: A decision logic language (DL-language) L is now introduced [1]. The set of attribute constants is defined as $a \in A$, and the set of attribute value constants is $V = \bigcup V_a$. The propositional variables are φ and ψ, and the propositional connectives are $\bot,\ \sim,\ \wedge,\ \vee\ and\ \rightarrow$.

Definition 4. The set of formulas of the decision logic language (DL-language) L is the smallest set satisfying the following conditions:

1) (a, v), or in short a_v, is an atomic formula of L.
2) If $\varphi\ and\ \psi$ are formulas of the DL-language, then $\sim \varphi,\ \varphi \wedge \psi,\ \varphi \vee \psi\ and\ \varphi \rightarrow \psi$ are formulas.

The interpretation of the DL-language L is performed using the universe U in $S = (U, A)$ of the Knowledge Representation System ($KR-system$) and the assignment function, mapping from U to objects of formulas. Formulas of the DL-language are interpreted as subsets of objects consisting of a value v and an attribute a.

Atomic formulas (a, v) describe objects that have a value v for the attribute a. $S \models_s \varphi$ denotes that the object $x \in U$ satisfies the formula φ of $S = (U, A)$. The semantics of DL-language is defined as follows:

$S \models_s (a, v)$ *iff* $a(x) = v$
$S \models_s \sim \varphi$ *iff* $S \not\models_s \varphi$
$S \models_s \varphi \vee \psi$ *iff* $S \models_s \varphi$ *or* $S \models_s \psi$
$S \models_s \varphi \wedge \psi$ *iff* $S \models_s \varphi$ *and* $S \models_s \psi$
$S \models_s \varphi \rightarrow \psi$ *iff* $S \models_s \sim \varphi \vee \psi$

Let φ be an atomic formula of the DL-language, $R \in C \cup D$ an equivalence relation, and X any subset of U. Then, the truth value of φ is defined as follows:

$$||\varphi||_s = \begin{cases} \mathbf{t}\ if\ |\varphi|_s \subseteq POS_R(U/X) \\ \mathbf{f}\ if\ |\varphi|_s \subseteq NEG_R(U/X) \end{cases}$$

This shows that decision logic is based on bivalent logic. In the next section, an interpretation of decision logic based on three-valued logics will be discussed.

III. Relationship with Three-valued Semantics

Partial semantics for classical logic has been studied by van Benthem in the context of the *semantic tableaux* [7][8]. In this section, the application of partial semantics to decision logic is investigated. As the proposed approach can replace the base (bivalent) logic of decision logic, alternative versions of decision logic based on three-valued logics are obtained.

The model $\mathcal{S}$ of decision logic based on three-valued semantics consists of a universe U for the language L and an assignment function s that provides an interpretation for L.

For the domain $|\mathcal{S}|$ of the model $\mathcal{S}$, a subset is defined by $S = \langle S^+, S^- \rangle$. The first term of the ordered pair denotes the set of n-tuples of elements of the universe that $verify$ the relation S, whereas the second term denotes the set of n-tuples that $falsify$ the relation. The interpretation of propositional variables of L for the model $\mathcal{S}$ is given by $S_{\mathcal{S}} = \langle (S)^+_{\mathcal{S}}, (S)^-_{\mathcal{S}} \rangle$. Let $T = \{t, f, u\}$ be the truth value for the three-valued semantics of L, where each value is defined as true, false, or undefined (or indeterminate). Then, the truth value of φ on $S = (U, A)$ is defined as follows:

$$||\varphi||_s = \begin{cases} \mathbf{t}\ if\ |\varphi|_s \subseteq POS_R(U/X) \\ \mathbf{f}\ if\ |\varphi|_s \subseteq NEG_R(U/X) \\ \mathbf{u}\ if\ |\varphi|_s \subseteq BN_R(U/X) \end{cases}$$

A semantic relation for the model $\mathcal{S}$ is defined following [7][9][10]. The truth and the falsehood of a formula of the DL-language are defined in a model $\mathcal{S}$. The truth (denoted by $\models^+_s$) and the falsehood (denoted by $\models^-_s$) of the formulas of the decision logic in $\mathcal{S}$ are defined inductively:

Definition 5. Semantic relation of $\mathcal{S} \models^+_s \varphi$ and $\mathcal{S} \models^-_s \varphi$ are defined as follows:

$\mathcal{S} \models^+_s \varphi$ *iff* $\varphi \in \mathcal{S}^+$
$\mathcal{S} \models^-_s \varphi$ *iff* $\varphi \in \mathcal{S}^-$
$\mathcal{S} \models^+_s \sim \varphi$ *iff* $\mathcal{S} \models^-_s \varphi$
$\mathcal{S} \models^-_s \sim \varphi$ *iff* $\mathcal{S} \models^+_s \varphi$
$\mathcal{S} \models^+_s \varphi \vee \psi$ *iff* $\mathcal{S} \models^+_s \varphi$ *or* $\mathcal{S} \models^+_s \psi$
$\mathcal{S} \models^-_s \varphi \vee \psi$ *iff* $\mathcal{S} \models^-_s \varphi$ *and* $\mathcal{S} \models^-_s \psi$
$\mathcal{S} \models^+_s \varphi \wedge \psi$ *iff* $\mathcal{S} \models^+_s \varphi$ *and* $\mathcal{S} \models^+_s \psi$
$\mathcal{S} \models^-_s \varphi \wedge \psi$ *iff* $\mathcal{S} \models^-_s \varphi$ *or* $\mathcal{S} \models^-_s \psi$
$\mathcal{S} \models^+_s \varphi \rightarrow \psi$ *iff* $\mathcal{S} \models^-_s \varphi$ *or* $\mathcal{S} \models^+_s \psi$
$\mathcal{S} \models^-_s \varphi \rightarrow \psi$ *iff* $\mathcal{S} \models^+_s \varphi$ *and* $\mathcal{S} \models^-_s \psi$

$\models^+$ denotes $confirmation$ and $\models^-$ *refutation*. The model $\mathcal{S}$ is *consistent* if and only if $S^+ \cap S^- = \emptyset$. The symbol $\sim$ denotes strong negation, in which $\sim$ is interpreted as true if the proposition is false.

Theorem 1. For every model $\mathcal{S}$, DL-language L, and formula φ, it is not the case that $\mathcal{S} \models^+_s \varphi$ and $\mathcal{S} \models^-_s \varphi$ hold.

 ISBN: 978-1-61208-600-2

Proof: Only the proof for $\sim$ and $\wedge$ will be provided. It can be carried out by induction on the complexity of the formula. The condition of *consistent* implies that it is not the case that $\varphi \in \mathcal{S}^+$ and $\varphi \in \mathcal{S}^-$. Then, it is not the case that $\mathcal{S} \models_s^+ \varphi$ and $\mathcal{S} \models_s^- \varphi$.

$\sim$: We assume that $\mathcal{S} \models_s^+ \sim\varphi$ and $\mathcal{S} \models_s^- \sim\varphi$ hold. Then, it follows that $\mathcal{S} \models_s^+ \varphi$ and $\mathcal{S} \models_s^- \varphi$. This is a contradiction.

$\wedge$: We assume that $\mathcal{S} \models_s^- \varphi \wedge \psi$ and $\mathcal{S} \models_s^+ \varphi \wedge \psi$ hold. Then, it follows that $\mathcal{S} \models_s^+ \varphi$ *and* $\mathcal{S} \models_s^+ \psi$ and $\mathcal{S} \models_s^- \varphi$ *or* $\mathcal{S} \models_s^- \psi$. This is a contradiction. ■

Example. We assume the decision table below, where the condition and decision attributes are not considered.

$U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$

Attribute: $C = \{c1, c2, c3, c4\}$

$c_1 = \{x_1, x_4, x_8\}, c_2 = \{x_2, x_5, x_7\}, c_3 = \{x_3\},$

$c_4 = \{x_6\}$

$U/C = c_1 \cup c_2 \cup c_3 \cup c_4$

Any subset $X = \{x_3, x_6, x_8\}$

$POS_C(X) = c_3 \cup c_4 = \{x_3, x_6\}$

$BN_C(X) = c_1 = \{x_1, x_4, x_8\}$

$NEG_C(X) = c_2 = \{x_2, x_5, x_7\}$

Evaluation of truth value of formulas as follows:

If $|C_{c3}| \subseteq POS_C(X)$ *then* $||C_{c3}||_s = \mathbf{t}$

If $|C_{c1}| \subseteq BN_C(X)$ *then* $||C_{c1}||_s = \mathbf{u}$

If $|C_{c2}| \subseteq NEG_C(X)$ *then* $||C_{c2}||_s = \mathbf{f}$

IV. Consequence relation and Sequent calculus

Partial semantics in classical logic is closely related to the interpretation of the Beth tableau [8]. Van Benthem [7] suggested the relationship of the consequence relation to Gentzen sequent calculus. Thus, the application of the consequence relation for partial semantics to decision logic will be discussed, as well as the structure of three-valued logic that is based on partial semantics and replaces the basic (bivalent) logic of decision logic.

To prove $X \to Y$ by the Beth tableau, a counterexample, such as $X \& \sim Y$, is constructed. Here, let X be Γ and Y be Δ (set of formulas), and let A and B be formulas.

Axiom: $A \Rightarrow A \ \ (ID)$

Sequent rule:

$$\frac{\Gamma \Rightarrow \Delta}{A, \Gamma \Rightarrow \Delta, A}\ (Weakening) \qquad \frac{\Gamma, A \Rightarrow \Delta \quad \Gamma \Rightarrow A, \Delta}{\Gamma \Rightarrow \Delta}\ (Cut)$$

$$\frac{A, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \sim A}\ (\sim R) \qquad \frac{\Gamma \Rightarrow \Delta, A}{\sim A, \Gamma \Rightarrow \Delta}\ (\sim L)$$

$$\frac{\Gamma \Rightarrow \Delta, A \quad \Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \wedge B}\ (\wedge R) \qquad \frac{A, B, \Gamma \Rightarrow \Delta}{A \wedge B, \Gamma \Rightarrow \Delta}\ (\wedge L)$$

$$\frac{\Gamma \Rightarrow \Delta, A, B}{\Gamma \Rightarrow \Delta, A \vee B}\ (\vee R) \qquad \frac{A, \Gamma \Rightarrow \Delta \quad B, \Gamma \Rightarrow \Delta}{A \vee B, \Gamma \Rightarrow \Delta}\ (\vee L)$$

$$\frac{A, \Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \to B}\ (\to R) \qquad \frac{\Gamma \Rightarrow \Delta, A \quad B, \Gamma \Rightarrow \Delta}{A \to B, \Gamma \Rightarrow \Delta}\ (\to L)$$

This axiomatization is based on the sequent calculus for classical logic LK (logistischer klassischer Kalkül) originally introduced by Gentzen in 1935 [11]. Decision logic is a predicate logic; however, in this study, the focus is on propositional logic without quantifiers and predicate symbols. This LK is extended to other deductive systems for partial semantics based on a different consequence relation. For example, the three-valued logic by Kleene has no tautology. Thus, to define a consequence relation, a logical system for three-valued logic is formalized. In the Beth tableau, to interpret the consequence relation for partial semantics, an atomic formula A with left open branch is evaluated as $V(A) = 1$, and an atomic formula B with right open branch as $V(B) = 0$. This can be interpreted according to sequent calculus. It is assumed that V is a partial assignment function assigning to an atomic formula the values 0 or 1. Then, the consequence relation is defined as follows:

(C1) *for all* V, *if* $V(Pre) = 1$ *then* $V(Cons) = 1$,

(C2) *for all* V, *if* $V(Pre) = 1$ *then* $V(Cons) \neq 0$.

Pre and $Cons$ represent sequent premise and conclusion, respectively. In classical logic, (C1) and (C2) can be interpreted as equivalent; however, they are not equivalent in partial logic based on partial semantics.

Sequent calculi G1 for (C1) can be obtained by adding the following rules to LK$\backslash\{(\sim R)\}$, where, "$\backslash$" implies that the rule following "$\backslash$" is excluded.

$$\frac{\Gamma \Rightarrow \Delta, A}{\Gamma \Rightarrow \Delta, \sim\sim A}\ (\sim\sim R) \qquad \frac{A, \Gamma \Rightarrow \Delta}{\sim\sim A, \Gamma \Rightarrow \Delta}\ (\sim\sim L)$$

$$\frac{\Gamma \Rightarrow \Delta, \sim A, \sim B}{\Gamma \Rightarrow \Delta, \sim (A \wedge B)}\ (\sim \wedge R)$$

$$\frac{\sim A, \Gamma \Rightarrow \Delta \quad \sim B, \Gamma \Rightarrow \Delta}{\sim (A \wedge B), \Gamma \Rightarrow \Delta}\ (\sim \wedge L)$$

$$\frac{\Gamma \Rightarrow \Delta, \sim A \quad \Gamma \Rightarrow \Delta, \sim B}{\Gamma \Rightarrow \Delta, \sim (A \vee B)}\ (\sim \vee R)$$

$$\frac{\sim A, \sim B, \Gamma \Rightarrow \Delta}{\sim (A \vee B), \Gamma \Rightarrow \Delta}\ (\sim \vee L)$$

These Gentzen-type sequent calculi axiomatize (C1) [12][7].

We are now in a position to define GC1. For GC1, (A1) defined below is added to G1$\backslash\{(\sim L)\}$.

(A1) $A, \sim A \Rightarrow$

GC1:=$\{(ID),\ (Weakening),\ (Cut),\ (A1),\ (\wedge R),\ (\wedge L),$
$(\vee R),\ (\vee L),\ (\to R),\ (\to L),\ (\sim\sim R),\ (\sim\sim L),$
$(\sim \wedge R),\ (\sim \wedge L),\ (\sim \vee R),\ (\sim \vee L)\}$

For the rule $(\sim L)$ obtained from (A1), GC1 and G1 are equivalent.

Theorem 2. GC1 = G1.

Proof: (A1) can be considered as $(\sim L)$, then double negation and de Morgan laws in GC1 are obtained. ■

The semantic relation of the implication of $\mathcal{S}$ for GC1 is defined in Definition 5.

Then, rule (C2) for the Gentzen system is axiomatized as GC2. GC2 is obtained by replacing axiom (A1) from GC1 to (A2) below.

(A2) $\Rightarrow A, \sim A$

By exclusion of the restriction in Theorem 1, the definition of the semantic relation for the implication of GC2 is obtained as follows:

$\mathcal{S} \models_s^+ \varphi \to \psi$ *iff* $\mathcal{S} \not\models_s^+ \varphi$ *or* $\mathcal{S} \not\models_s^- \psi$ or
$(\mathcal{S} \models_s^+ \varphi$ *and* $\mathcal{S} \models_s^- \varphi$ *and* $\mathcal{S} \models_s^+ \psi$ *and* $\mathcal{S} \models_s^- \psi)$

$\mathcal{S} \models_s^- \varphi \to \psi$ *iff* $\mathcal{S} \models_s^+ \varphi$ *and* $\mathcal{S} \models_s^- \psi$

Theorem 3. C2 is axiomatized by GC2.

Proof: GC2 is an axiomatization which is obtained from GC1 by replacing (A1) with (A2). ■

There are some possible options to define consequence relation. For our purposes, (C3) below is proposed as alternative definition.

(C3) *for all* V, *if* $V(Pre) = 1$ *then* $V(Cons) = 1$,
if $V(Cons) = 0$ *then* $V(Pre) = 0$.

The Gentzen system GC3 for (C3) is obtained by replacing (A1) of GC1 with the following (A3):

(A3) $A, \sim A \Rightarrow B, \sim B$

V. Relationship Partial Logic

In this section, the relationship between the sequent calculi system based on partial semantics and three-valued logic is discussed. The three-valued logic is extended by defining the weak negation $\neg$. $\sim$ is treated as the strong or classical negation. Weak negation represents the lack of truth. In partial semantics, it allows an interpretation whereby $\neg$ is true if a proposition is not true, that is false or undefined. The semantic relation for weak negation is as follows:

$\mathcal{S} \models_s^+ \neg\varphi$ *iff* $\mathcal{S} \not\models_s^+ \varphi$

$\mathcal{S} \models_s^- \neg\varphi$ *iff* $\mathcal{S} \models_s^+ \varphi$

The truth value of weak negation is defined as follows:

$$||\neg\varphi||_s = \begin{cases} \mathbf{t} \text{ if } ||\varphi||_s = \mathbf{f} \text{ or } \mathbf{u} \\ \mathbf{f} \text{ if } ||\varphi||_s = \mathbf{t} \end{cases}$$

By introducing weak negation, the representation of deduction for uncertain concepts may be handled; however, this is beyond the scope of this study. Moreover, weak implication may be defined using weak negation as follows:

$A \rightarrow_w B =_{def} \neg A \vee B$

The following rules for weak negation and weak implication are now presented.

$$\frac{\Gamma \Rightarrow A, \Delta}{\Gamma \Rightarrow \neg A, \Delta}\ (\neg R) \qquad \frac{\Gamma, A \Rightarrow \Delta}{\Gamma, \neg A \Rightarrow \Delta}\ (\neg L)$$

$$\frac{A, \Gamma \Rightarrow \Delta, B}{\Gamma \Rightarrow \Delta, A \rightarrow_w B}\ (\rightarrow_w R) \qquad \frac{B, \Gamma \Rightarrow \Delta \quad \Gamma \Rightarrow \Delta, A}{A \rightarrow_w B, \Gamma \Rightarrow \Delta}\ (\rightarrow_w)$$

Three extended decision logics (EDLs) based on three-valued logic are subsequently presented. They are adapted to handle ambiguity and uncertainty. GC1, which was discussed above, is interpreted as a strong Kleene three-valued logic. It is first assumed that GC1 is the basic deduction system for decision logic. Then, the inference rules of weak negation and weak implication are added. This logic is the extended decision logic EDL1. Its semantic relation is denoted by $\models_{EDL1}$.

The axioms and rules of EDL1 are as follows:

EDL1 := $\{(ID), (Weakening), (Cut), (A1), (\wedge R), (\wedge L),$
$(\vee R),\ (\vee L),\ (\rightarrow R),\ (\rightarrow L),\ (\sim\sim R),\ (\sim\sim L),$
$(\sim \wedge R),\ (\sim \wedge L),\ (\sim \vee R),\ (\sim \vee L),$
$(\neg R),\ (\neg L),\ (\rightarrow_w R),\ (\rightarrow_w L)\}$

The concept of a proposition that is neither true nor false is possible in EDL1. If the designated value of three-valued logic of GC2 is defined as {t, u}, then this system is a paraconsistent logic. Paraconsistent logic does not hold for the principle of explosion (ex falso quodlibet); therefore, it is possible to interpret the consequence relation by (C2). The semantic relation of EDL2 is obtained from EDL1 by replacing (A1) with (A2).

EDL2 := EDL1\{(A1)} + {(A2)}

The semantic relation of EDL3 is obtained from EDL1 replacing (A1) with (A3).

EDL3 := EDL1\{(A1)} + {(A3)}

EDL3 is interpreted as both paracomplete and paraconsistent. This prevents the paradox of material implication of classical logic. In decision logic, the decision rule is interpreted as follows: If the premise is valid, then the conclusion is also valid. If the conclusion is not valid, then the premise is not valid either.

VI. Conclusion and Future Work

It was proposed that a partial semantics interpretation of the consequence relation may serve as a foundation for decision logic. A three-valued logic system based on a consequence relation that is defined by partial semantics was investigated, and the relationship between them was studied. By adopting three-valued logic as basic logic for decision logic, its deductive system can be enhanced. Moreover, this allows the extension of the scope of its application.

In future work, the semantic relationship between decision logic and partial semantics should be investigated in detail. Furthermore, soundness and completeness results should be derived for extended decision logic. This is required for the foundation of a logical system for decision logic. Finally, the application of decision logic based on three-valued logic should be investigated.

Acknowledgment

We thank anonymous referees for valuable feedback on an earlier version of this paper.

References

[1] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data," Kluwer Academic Publishers, 1991.

[2] A. Avron and B. Konikowska, "Rough Sets and 3-Valued Logics," Studia Logica, vol. 90, 2008, pp. 69–92.

[3] D. Ciucci and D. Dubois, "Three-Valued Logics, Uncertainty Management and Rough Sets," in Transactions on Rough Sets XVII, Lecture Notes in Computer Science book series (LNCS, volume 8375), 2001, pp. 1–32.

[4] A. Urquhart, "Basic Many-Valued Logic," Handbook of Philosophical Logic, vol. 2, 2001, pp. 249–295.

[5] T.-F. Fan, W.-C. Hu, and C.-J. Liau, "Decision logics for knowledge representation in data mining," in 25th Annual International Computer Software and Applications Conference. COMPSAC, 2001, pp. 626–631.

[6] Y. Lin and L. Qing, "A Logical Method of Formalization for Granular Computing," IEEE International Conference on Granular Computing (GRC 2007), 2007, pp. 22–22.

[7] J. Van Benthem, "Partiality and Nonmonotonicity in Classical Logic," Logique et Analyse, vol. 29, 1986, pp. 225–247.

[8] R. Smullyan, "First-Order Logic," Dover Books, 1995.

[9] V. Degauquier, "Partial and paraconsistent three-valued logics," Logic and Logical Philosophy, vol. 25, 2016, pp. 143–171.

[10] R. Muskens, "On Partial and Paraconsistent Logics," Notre Dame J. Formal Logic, vol. 40, 1999, pp. 352–374.

[11] G. Gentzen, "Untersuchungen über das logische Schliesen. I," in Mathematische Zeitschrift, vol. 39. Springer-Verlag, 1935, pp. 176–210.

[12] S. Akama and Y. Nakayama, "Consequence relations in DRT," Proc. of The 15th International Conference on Computational Linguistics COLING 1994, vol. 2, 1994, pp. 1114–1117.

 ISBN: 978-1-61208-600-2

Towards the Development of a CTS2-based Terminology Service in the Italian Federated Electronic Health Record

Elena Cardillo and Maria Teresa Chiaravalloti

Institute of Informatics and Telematics
National Research Council, IIT - CNR
Rende, Italy
Email: elena.cardillo@iit.cnr.it, maria.chiaravalloti@iit.cnr.it

***Abstract*— Semantic interoperability is essential for advanced Electronic Health Records (EHRs) functionality, and in particular for data exchanges, and efficient communication among clinicians. Integrated terminology services offer the chance to manage clinical code systems, both standard and local, and value sets, through a series of functionalities such as searching, querying, cross mapping, etc. The main standard in the domain is Clinical Terminology Service Release 2 (CTS2) by Health Level 7 (HL7). This paper describes the approach used for designing and developing an integrated terminology service based on the CTS2 standard, namely *Servizio Terminologico Integrato* (STI), which aims to support domain experts and healthcare organizations in ensuring semantic interoperability in the Italian Federated EHR.**

Keywords—Coding Systems; Semantic Interoperability; Terminology Services; CTS2; Healthcare.

I. INTRODUCTION

Interoperability of clinical data essentially means that different systems are able to communicate among each other, exchange data, and, above all, reuse them. The general aim is ensuring a worldwide availability of information at the right time and place, in order to deliver better clinical services and improve healthcare. Interoperability is a required function for the proper use of Electronic Health Record (EHR) systems, which remain simple data containers if they do not have the chance to communicate using the same language. Standardized coding systems are the lingua franca of medical data, as they allow to uniquely identify the same concept despite languages, synonyms and local names that could be used to refer them. The advantages of standards are commonly recognized and their usefulness increases over time as they are employed in numerous health-related Information Technology applications. Nonetheless, their use is not always as easy as it may appear and health professionals often complain about the lack of adequate support systems.

Managing clinical terminologies is not only a matter of making them available to the users, but their management needs to include further functions to offer a complete plethora of services allowing a meaningful use of standards. To pursue this aim, there is the need of a standard protocol to manage terminology standards in the same way across multiple healthcare facilities. This role is covered by integrated terminology services, which offer the possibility to interact with terminologies according to a series of standardized functionalities, such as research, hierarchical tree navigation, structured query, cross mapping.

The Italian National Research Council (CNR) is working in accordance with the Digital Italy Agency (AgID) for realizing the national federate EHR (corresponding to the Italian acronym FSE, meaning "Fascicolo Sanitario Elettronico") infrastructure in order to allow the exchange of clinical documents among the regional EHR systems [1]. In this frame, semantic interoperability is a non-trivial issue [2], especially because, over time, regional and local coding systems and habits have proliferated. The Prime Minister Decree No. 178/2015 [3] disciplines the use of the FSE and makes some medical terminologies mandatory, detailing their use in the two kinds of documents (Patient Summary and Laboratory Report) included in the minimum unit expected in the FSE first implementation phase.

The objective of this paper is to describe the approach used for designing and developing an integrated terminology service, called STI (acronym of the Italian *Servizio Terminologico Integrato*), to support Regions, domain experts and health facilities in the management of the clinical coding systems and terminologies prescribed by the cited Decree, and to ease their use in the documents required in the FSE. To develop this terminology service, the standard protocol HL7 Common Terminology Services Rel. 2 (CTS2) [4] was tested.

The paper is organized as follows. Section II gives an overview of the state of the art on semantic interoperability and the main features of CTS2. Section III describes the material used within the Italian implementation of the STI system and Section IV addresses the content approach. Section V shows some preliminary results. Discussion and conclusions in Section VI close the paper.

II. BACKGROUND

A. Semantic Interoperability: projects and initiatives

The adjective *semantic* conveys the deep meaning of interoperability as it overcomes lexical and syntactical issues to deal with the meaning of the exchanged information. The best EHR system would be useless without semantic interoperability, as it could not unambiguously interpret data received from other systems. In fact, Semantic Interoperability in healthcare is defined as "the ability of a

 ISBN: 978-1-61208-600-2

healthcare system to share information and have that information properly interpreted by the receiving system in the same sense as intended by the transmitting system" [5].

Projects and initiatives address the semantic interoperability issue trying to propose effective solutions to solve it. Regarding European Community (EU) initiatives and projects, it is worth mentioning the FP7 project *Semantic Interoperability for Health Network*, whose main aim was the implementation of the necessary infrastructure and governance to allow a sustainable semantic interoperability of clinical and biomedical knowledge at European level [6]. Furthermore, the project EHR4CR [7] dealt with the development of a semantic interoperability service platform, which includes a mediation model for multiple standards integration and harmonization. It was tested in 11 EHR systems of 5 EU Countries. Finally, the Trillium Bridge and Trillium Bridge II projects involve EU Countries and US for the creation of a shared model of an International Patient Summary (IPS), to improve semantic interoperability of e-health systems beyond EU borders [8].

Also, international standards organizations proposed protocols for semantic interoperability. The main one is the CTS2 standard proposed in the Healthcare Service Specification Program (HSSP), a joint HL7 and Object Management Group (OMG) initiative [9]. CTS2 is a cohesive model and specification for representing, accessing, querying, exchanging and updating terminological resources (e.g., Code Systems, Value Sets, Mappings), built on the RESTful (Representational state transfer) Architectural Style. More recently, HL7 proposed the Fast Healthcare Interoperability Resources (FHIR) Specification [10], another standard for exchanging healthcare information electronically, which, compared to the previous HL7 standards, is more consistent and easy to implement, thanks to its built-in extension mechanism to cover the needed content. In fact, specific use cases can be implemented by combining resources together through the use of resource references.

In the literature, different initiatives aimed at developing terminology integration platforms or services were launched. Initial studies and applications focused, for example, on the use of the Unified Medical Language System (UMLS) Metathesaurus, developed by the US National Library of Medicine [11], which includes more than 100 biomedical vocabularies integrated on the basis of a common Semantic Network and mapped among them. Researchers used UMLS to create knowledge-based representation for controlled terminologies of clinical information and to extract and validate semantic relationships. Particularly relevant are also the HETOP terminology service, which includes cross lingual multi-terminological mappings on a semantic basis [12]; and the LexGrid initiative [13], which promotes the use of common terminology models to accommodate multiple vocabulary and ontology distribution formats, as well as the support of multiple data storing for federated vocabulary distribution.

In the last few years, much effort has been spent on the application of the abovementioned CTS2 standard to develop terminology services, as that realized by the Mayo Clinic Informatics, which is the most internationally relevant [9]. D2Refine Workbench platform, for example, aimed at standardizing and harmonizing clinical study data dictionaries [14]. Focused on laboratory catalogues, the experience of the Partners HealthCare System of Boston, applies the CTS2 *Upper Level Class Model* to represent and harmonize the structure of both local laboratory order dictionaries and reference terminologies [15]. Peterson et al. presented, instead, a design user-centered approach, based on the use of Extraction, Transformation and Loading (ETL) procedures in CTS2-based terminology services [16]. The main advantages of this service are: i) adaptability, ii) interoperability, because of the numerous standard vocabularies included, iii) usability, since focused on users' needs. In the wake of these projects, we propose a multi-layers CTS2 implementation that is not only based on ETL procedures, but allows also mapping (and their validation) between local dictionaries and standardized code systems, including also semantic enrichment through external ontological references.

Interesting applications of CTS2 can be found also in the European context, where the main implementation is the Standard Terminology Services (STS) provided by the French non-profit development standards and services organization PHAST [17]. Other implementations are the following: the Austrian national patient health record ELGA, where, all relevant clinical terminologies are provided through a CTS2-conformant terminology server [18]; and the Terminology Server, realized by the University of Applied Science of Dortmund, which also offers a collaboration environment to develop terminologies in a team [19]. Finally, in Italy, the existing implementations of CTS2-based terminology services are proprietary solutions, i.e., the Distributed Terminology Assets Management system (DITAM) [20]; and the HQuantum [21], which is especially focused on the management and integration of local laboratory data through the LOINC standard. These two solutions were evaluated as non-fitting for the purpose of our project because they are subject to license, while the FSE project required an open and reusable solution. Furthermore, they were, at that time, only partially developed and tested, and this would have implied a lot of customization effort.

B. CTS2 HL7 overview

As stated in the ANSI/HL7 V3 CTS R2-2015 standard [4], "the HL7 Common Terminology Services (HL7 CTS) is an API specification that is intended to describe the basic functionalities that will be needed by HL7 Version 3 software implementations to query and access terminological content. It is specified as an Application Programming Interface (API) rather than a set of data structures to enable a wide variety of terminological content to be integrated within the HL7 Version 3 messaging framework without the need

for significant migration or rewrite". The standard, currently, consists of:

- CTS2 Normative Edition v1.0 and the Service Functional Model (SFM), that serve as Functional Requirements Documents, defining the capabilities, responsibilities, inputs, outputs, expected behavior and a set of core functionalities to support the management, maintenance, and interaction with ontologies and medical vocabulary systems.
- CTS2 Technical Specification, that serves as a technical specification document to define the precise API interface specifications for CTS2 implementation compliance in Simple Object Access Protocol (SOAP).

The CTS2 Information Model specifies the structural definition, attributes and associations of Resources common to structured terminologies such as Code Systems, Binding Domains and Value Sets. The Computational Model specifies the service descriptions and interfaces needed to access and maintain structured terminologies. The main CTS2 profiles and functionalities are:

a) *Search/Query Profile*, including: reading of a resource, a code or a concept; browsing or visualization of the tree of a resource; the download of a resource.
b) *Terminology Administration Profile*, including administration functionalities: import of a resource; creation of mappings between the imported resources; the possibility to use updates and notifications.
c) *Terminology Authoring Profile*, including "read-write" functionalities intended for an application used by specialized users (e.g., translators) to create and maintain terminological resources.

More specifically:

- *Read* – the direct access to the resource content via URI, local identifier or a combination of an abstract resource identifier and version tag (e.g., LOINC/Current version).
- *Query* – the ability to access, combine and filter lists of resources based on their content and user context.
- *Import and Export* – the ability to import external content and/or export the contents of the service in different formats.
- *Update* – the ability to validate load sets of changes into the service that updates its content.
- *History* – the ability to determine what changes occurred over stated periods of time.
- *Maintenance* – the ability to create and commit sets of changes.
- *Temporal* – the ability to query on the state of the service at a given point in the past (or in the future).
- *Specialized* – service specific functions such as the association reasoning services, the map entry services and the resolved value set services.

The CTS2 Development Framework (DF) is a development kit for rapidly creating CTS2 compliant applications. It allows users to create plugins, which may be loaded into the DF to provide REST Web Services that use CTS2 compliant paths and model objects. Since it is plugin based, users are only required to implement the functionality that is exclusive to their environment. Thus, CTS2 DF provides all the infrastructures and utilities to help users create plugins. Given the short time available to develop the service, in this work, we reused the mentioned CTS2 DF toolkit provided by Mayo Clinic Informatics [9], which is useful for rapidly creating CTS2 compliant applications, and, at the moment, it is recognized as the most complete and documented. Furthermore, the community that uses the DF is wide and quite reactive.

III. MATERIAL

The terminology service was designed taking into account both the CTS2 main functionalities requirements and the structures of the medical coding systems required by the law (described in Table I).

TABLE I CODE SYSTEMS REQUIRED IN FSE

Code Systems	Maintaining Organizations	Use in FSE
Internatioanl Classification of Disease 9th Revision, Clinical Modifications (ICD-9-CM)	World Health Organization	Required in the Patient Summary for coding relevant and chronic diseases, in Prescriptions and in Discharge Letters for coding Diagnoses
Anatomical Therapeutic Chemical Classification (ATC)	WHO Collaborating Centre for Drug Statistics Methodology Norwegian Institute of Public Health	Required in the Patient Summary for coding adverse reactions to food and medication, medication plan, and vaccinations
Autorizzazione all'Immissione in Commercio (AIC)	Italian Medicines Agency	Required in the Patient Summary for coding medications
Logical Observation Identifiers Names and Codes (LOINC)	Regenstrief Institute	Required in the Laboratory Report for coding performed tests and their specialty or class

As each resource has a different structure, the most suitable solution was integrating them into the STI Knowledge Base (KB) allowing the correct visualization and searching into each of them. ICD-9-CM, for example, is a classification, which has a hierarchical tree structure so in the visualization it needed the use of indentations and expandable/collapsible branches for navigating the tree; LOINC, instead, is more like a nomenclature, without any hierarchical structure (codes are progressive and not informative). Furthermore, each LOINC code has associated numerous information to be visualized, which are discriminative in choosing a code rather than another, so it

 ISBN: 978-1-61208-600-2

was necessary to realize a personalized form to access LOINC code details (e.g., system, scale, method, etc.).

As further explained in Section IV.B, a great deal of effort was spent on the collection of the different versions of each standard and on re-structuring the available files according to the CTS2 concept model, to integrate them in the KB. Other than the standard terminologies required by the law, some other resources, such as value sets and local files mapped to the code systems, were integrated into STI. The first type includes files of synonyms of LOINC and ICD-9-CM terms, which could be used as further research items to find a specific code. Local files mapped to the code systems are useful in STI as both basis of comparison for who is working on the same type of mapping and collector of local synonyms of the official terms used in the standard clinical terminologies. At the moment, a file of LOINC mapped local tests of some laboratories of the Umbria Region is in service.

IV. APPROACH

The proposed solution consists of a standard-based and web-based distributed software infrastructure, which is open and extendable, and aims to support the production, integration, maintenance, and use of the terminological resources according to the CTS2 protocol. To design and develop the terminology service, an *Agile* methodology was applied. This led to an iterative development of the system functionalities, starting from the core ones and continuing with further iterations in the process of analysis and development. Each iteration and progress in the design and development of the functionalities were submitted to tests by terminology and domain experts.

A. STI Architecture

The STI Architecture (see Figure 1), was designed and realized by using Full Open Source integrated components:

- The Content Management System Liferay 6.2. CE [22] as environment to create the Web Application. It manages simultaneous user accesses, content versioning and classification. The platform functionalities were realized through the development of appropriate portlet allowing the management of: i) search and visualization of code systems; ii) administration management of import and elimination of code systems.
- Kettle (Pentaho Data Integration) [23], used to realize ETL procedures for data integration during data migration from different Database Management Systems. ETL procedures include: (i) heterogeneous data aggregation; (ii) data transport and transformation, by performing data cleaning operations, or scheduled-based data storing, on the destination database. ETL procedures are mostly used in the construction and population of the KB.
- Virtuoso Open Source Edition [24], developed by OpenLink, used for the management of ontologies and data in RDF. RDF data can be queried through SPARQL endpoints, to facilitate the connection with structured dataset derived from other sources.
- CKAN [25], for the management and publication of Open Data. This open source software allows cataloging datasets and describe them across a range of metadata that, on the one hand, help users to navigate through information, and on the other hand, facilitate indexing of the same datasets on search engines. In the present work CKAN is useful to export data (i.e., resources in STI) in the Open Data format and to publish them on open data platforms.

The strengths of this architecture and implementation are various: all the components are open source; it is scalable, modular and easy to maintain; it is installable on open environment without the need for a license.

B. The STI Knowledge Base

The implementation of the STI KB started with the integration of the basic elements, represented by the code systems. They were processed by ETL procedures, in order to enrich them with knowledge derived from external services. The modeling of the basic entities contained in these code systems was made through Porting on the Database. The definition of the STI KB was based on four application layers: 1) the Data layer; 2) the Integration layer; 3) the Semantic layer; and 4) the Presentation layer.

1) The Data layer: it is represented by the CTS2 Conceptual Model, for the representation of the different types of resources and semantic relationships, and by a relational DB, containing the useful facilities to integrate the resources, in particular code systems, in compliance with CTS2. Each concept of the code systems represents the basic entity that composes the atomic information of the conceptual model and is classified according to the structure defined in the HL7 standard, in XML format. In the STI KB, the code systems described in Section III are included in different versions after a readaptation of their structure to the CTS2 model, but maintaining, at the same time, their specifications. In particular, the collected versions are:

- ICD-9-CM Italian 2007 version, counting more than 16,000 codes. Since the official CSV file distributed on the Ministry of Health website is incomplete (it includes only code/description pairs), it was necessary to integrate it. To this aim, we reused an ontological version built in another project, where each ICD-9-CM code has different attributes: i) description, ii) alternative descriptions, i.e., synonyms, iii) inclusions, iv) exclusions, v) information on primary diagnoses, vi) information on additional diagnoses, vii) further notes.
- LOINC Italian and English versions 2.34 (required by the cited Decree, which counts more than 43,152 terms in the Italian version), 2.52, 2.54, 2.56 and 2.58 (which includes the latest updates of the system, released in December 2016, see Table II for details about LOINC terms). In order to correctly integrate LOINC in the STI KB, we needed to upload, for each version, three CSV files:

 ISBN: 978-1-61208-600-2

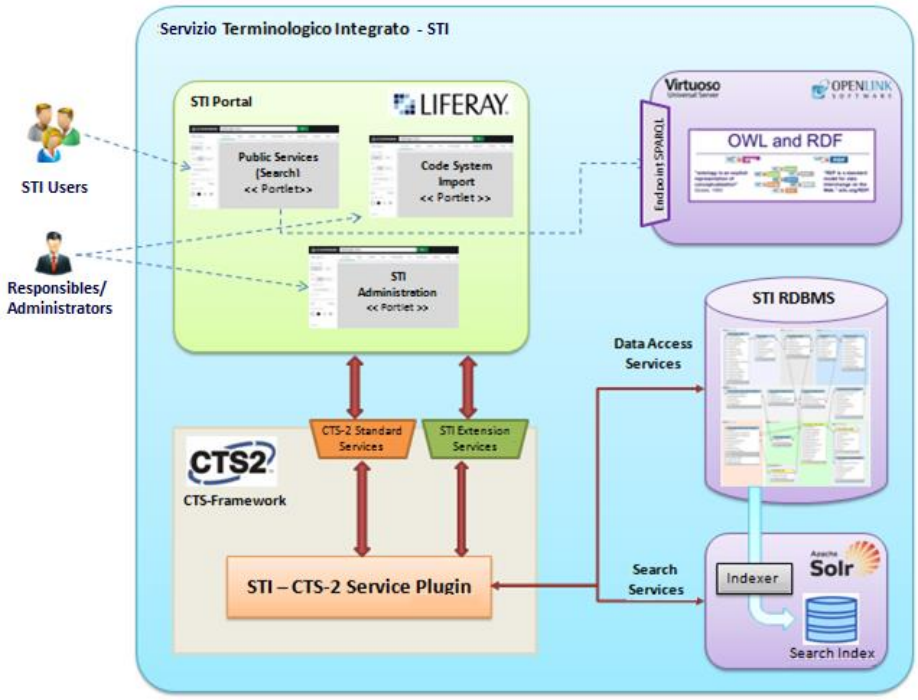

Figure 1. STI Architecture – Deployment Diagram

i) the Italian DB, LOINC_IT (which has limited number of fields associated to each code, with respect to the English version); ii) the English database, *LOINC_DB* (whose structure and fields changed several times during the updates), and iii) the file including the changes of the mapping codes from one version to another, named *Map_to*. For the LOINC_DB, it was necessary to make all the versions compliant with the structure of the last updates (v. 2.58) and to align the CSV of the Italian version to the same structure.

- AIC January 2017 version (the latest available updates on the AIFA website [26] at that moment, including more than 18,000 medicines codes). More specifically, AIC related files are published on the AIFA website as separate files according to the type of drugs. In particular, there are four different csv files: i) *equivalent_medicines* file, which includes for each AIC code the mapping to the active ingredient and thus the corresponding ATC code; ii) *Class_A_medicines*, including essential medicines and those for chronic illnesses (some of them can appear also in the equivalent medicines file); iii) *Class_H_medicines*, including medicines used only in hospital facilities, which can therefore not be sold in public pharmacies (some of them can appear also in the equivalent medicines file); and iv) *Class_C_medicines*, including drugs that are not licensed by the National Health Service (namely SSN) and are therefore paid by the citizen (all AIC codes in this file are mapped to the corresponding ATC code). These four files were separately integrated into the STI KB and ETL processes were used to clean and normalize data, in order to avoid concepts overlap.
- ATC Italian 2014 version (the latest one freely available at that moment), which counts about 5,000 codes. As for ICD-9-CM, access and navigation of the ATC classification tree was provided.

As AIC and ATC cover the same semantic area, a cross mapping file is constantly updated and available, but not in the form of a unique CSV file with 1-1 mappings. As mentioned above, in fact, only two of the four collected files contain mappings to ATC. In order to create a complete 1-1 mappings file, ETL processes were trained to perform mapping extraction processes from the data into the KB. In particular, where an explicit mapping in the different CSV files was not available, it was created by matching the active ingredient of the medicine and by querying external resources (i.e., the AIFA drugs database) in case of ambiguities (i.e., multiple AIC codes associated to one active ingredient), the status changes in AIC, and ATC code updates (considered that we did not use in our KB the latest version of the system). The final mapping file establishes mappings from one ATC code to multiple AIC codes, in fact AIC uniquely identifies branded medications while ATC encodes the medication active ingredient. This file was stored as a Mapping resource into the STI in order to give the chance to have a cross reference between the two code systems. In the Semantic layer, the basic information

 ISBN: 978-1-61208-600-2

included in the code systems is enriched by semantic content and correlations derived from the Integration layer.

Table II gives some statistics about the content integrated in STI Knowledge Base.

TABLE II STI KB STATISTICS

Resources	*Version*	*N. of Concepts (En)*	*N. of Concepts (It)*
LOINC	2.34	+ 60,000	43,152
LOINC	2.52	+ 72,000	58,045
LOINC	2.54	+ 73,000	61,419
LOINC	2.56	~ 80,000	60,837
LOINC	2.58	+ 80,000	63,367
ICD-9-CM	2007	16,100	16,100
ATC	2014	-	5,530
AIC	January 2017	-	18,309
LOINC 2.54 – Umbria_laboratory tests Catalogue Mapping	2016	-	111
AIC – ATC Mapping	2017	-	18,309
Total		+ 400,000	345,179

As can be seen, not all the code systems are mapped, except for *AIC - ATC*, and the *Umbria laboratory catalogue - LOINC*. Since we have not found official and available mappings between the other resources, we provided, as explained in Section V, a functionality to edit mappings directly on the STI platform, under specific permissions and subjected to validation by special users.

2) The Integration layer: it is used for the phase of design and modelling of the data transformation process, semantic enrichment by means of external endpoints, and for the internal organization of data in the KB. To this aim, we used the Kettle component and ETL procedures. In STI, ETL is a key process to bring all the data derived from code systems, which are heterogeneous in their structures and formats, in a standard, homogeneous environment. In particular, during the transformation phase, imported code systems are manipulated to be compatible with the target system (CTS2 model). In some cases, the necessary transformation rules are trivial, but in other situations (as happened for AIC and LOINC files, which changed database structures over the different updates) it may be necessary to sort, unite, and aggregate data. Pentaho Kettle provides a wizard that guides in the migration process, defining the source database server, the destination one, the mapping of the data types, and so much, so that migration does not cause data loss.

The population of the KB can be:

- manual, through the Web Application interface, by means of the compilation of specific forms and the selection of the relationships and classifications in the KB;
- through Rest services. The system will allow to import resources within the KB;
- semi-automatic, through the use of specific ETL systems that guarantee the information extraction and enrichment by querying external services, and finally adding the data to modelled knowledge.

3) The Semantic layer: it is based on the use of ontologies to extend data related to the atomic units (concepts in STI) with external components that have the related knowledge (e.g., the ontologies related to LOINC or ICD-9-CM concepts available in Bioportal [27]). To this aim, the Virtuoso platform was used. In particular, for each LOINC, ATC, or ICD-9-CM concept, it is possibile to query the Virtuoso platform in order to retrieve additional semantic/ontological information (e.g. the semantic type class of the LOINC concept *Hemoglobin A*, code "45208-6" is *Amino Acid, Peptide, or Protein*, or for example the LOINC concept *Aciclovir*, code "1-8", is a pharmaceutical substance whose semantic type is *Nucleic Acid, Nucleoside, or Nucleotide*).

4) The Presentation layer: it is the interface that uses the KB, characterized by the conceptual entities and enriched by a series of information. In particular, each concept in the KB was enriched by the following information provided by using different panels in the interface:

- code details, including information derived from the structure of the code system;
- status and versioning, including information on the changes of the status of a code and on the different versions available in the system;
- relationships, including the relationships of a precise concept to other concepts in the code systems (e.g., the hierarchical relationships of the three digit ICD-9-CM code 282 *Hereditary hemolytic anemias* to its leaf codes, etc.);
- mapping, where all the mappings of the selected code/concept to other resources available in the STI KB are visualized, if present;
- HL7 Specifications, including information needed to exchange data according to HL7 standard, i.e., Code, Code System OID, Code System Name, Code System Version, and concept Display Name;
- ontology, which gives access to the components LodView, to visualize the RDF data of a concept, and LodLive, to navigate the graph of a concept in the ontology derived from Bioportal.

An important aspect of the STI KB is that all the resources, where applicable, were imported in bilingual versions. In particular, LOINC, ATC and ICD-9-CM are available in English and in their official Italian version. The Italian version is in most cases aligned to the corresponding English one. Exceptions are LOINC, where the Italian

 ISBN: 978-1-61208-600-2

translation, as all other international LOINC translations, is always aligned to the previous LOINC English version; and the Italian translation of ATC, available in the system as version 2014, since it was not possible to collect the last Italian updates by the responsible government agencies.

C. Web services development

Considering the CTS2 functionalities described in Section II.A, we selected and implemented the following services in STI:

- *Reading*: reads the list of resources in STI and shows complete information on a single resource;
- *Search*: allows to search the resources for keywords or in particular fields thanks to the application of personalized filters;
- *Import*: allows to import the resources into the KB; the dataset within the CKAN component; rdf/owl graphs into the Virtuoso triple store;
- *Export*: allows to export complete resources in CSV or JSON formats;
- *Update*: allows the editing of the KB content;
- *Mapping*: allows the visualization of the existing mappings or the editing of new cross-mappings between the resources in STI;
- *Editorial workflow*: allows the approval of a particular resource or change in the KB, as well as the validation of new mappings created by users with special permissions.

The listed services are to be used through specific Rest Services useful for the reuse of the STI functionalities and for the interoperability with other systems. They will allow interoperability between the systems Liferay-CKAN-Virtuoso. For the development of these services, the Spring Web MVC framework was used. It provides Model-View-Controller (MVC) architecture and ready components that can be used to develop flexible and loosely coupled Web applications. The MVC pattern results in separating the different aspects of the application (input logic, business logic, and UI logic), while providing a loose coupling between these elements.

In a Linked Data perspective, STI service allows semantic enrichment of a resource, thus obtaining relationships with related resources, by querying exposed SPARQL endpoint, as the Bioportal ones.

D. Web Application Development

Regarding the Web Application, the system covers the following functionalities:

- User registration, authentication, roles and permissions management.
- CMS management (platform browsing management, content and versioning management, etc.).
- Multilingualism/bilingualism management (possibility to switch from Italian to English language when browsing a resource).
- Resource utilization by means of the Reading Web service.
- Search of one or more resources or concepts by means of the Search Web service.
- Resource and workflow management.
- Import and Export.
- Mapping between resources.
- Browsing of ontological resources by means of LodLive and LodView, linked to the Virtuoso SPARQL endpoint.
- Use of SPARQL endpoint for the resources stored in Virtuoso. Virtuoso will expose the resources in the RDF/TTL format via the SPARQL endpoint, allowing users to make more sophisticated queries.
- Management of the STI dataset imported in CKAN.

The Web Application allows to navigate the available resources according to the type (i.e., Code System, Value Set, Mapping). After the selection of a specific resource, depending on the original structure, it is possible to navigate the hierarchical tree, to directly select a code and visualize its details; or to search on the selected code system by using filters or full text search. The interface of the navigation functionality was built taking inspiration by the cited *Terminology Server*, the CTS2 implementation provided by the abovementioned University of Applied Sciences and Arts Dortmund.

V. RESULTS AND EVALUATION

STI service was released in its beta version in April 2017, as both Web service and Web application. It contains four standard code systems, which are those prescribed by the Law Decree regarding FSE, in their Italian and English versions, and also some mapping resources (local mapping to LOINC and AIC-ATC mappings). They can be accessed through the CTS2 main functionalities, such as searching, querying, navigation. Versions available are both those fixed by the cited law (i.e., LOINC 2.34 – December 2010) and the most recent ones (i.e., LOINC 2.58 – December 2016), so users can choose which one best fits their needs.

The service is open to the possibility of uploading additional code systems, mapping and value sets. They will be integrated, as it was for the four standards already available in the STI, taking into account their peculiar structure so to ensure a proper use of them. Furthermore, more local files mapped to the standard code systems can be uploaded by the system administrators after validation of their correctness. Regarding the mapping, there is also the chance, for users with special permissions (e.g. physicians, laboratory technicians, etc.), to create mappings between the available resources directly through the STI platform by using the *Cross-Mapping* functionality. During the cross-mapping, users have to qualify the mapping that they are creating between two concepts belonging to two different code systems, by selecting the type of association between the two selected concepts (e.g., choosing if two concepts are synonyms, clinically correlated, or if one is the hypernym of the other, etc. These cross-mappings, in any case, will be validated by the system administrators before becoming effective and saved in the STI KB. Regarding the interoperability services, as said in Section IV, STI allows

 ISBN: 978-1-61208-600-2

external applications, e.g., other terminology services installed at a regional level, to make requests to the Web service, which are those provided by the CTS2. In particular, the following examples are given:

1. Entity Description Query Service

Example: Search the entity *Immunoglobulina* in the code systems ICD9-CM:

- http://sti.iit.cnr.it/cts2framework/entities?matchvalue=immunoglobulina&page=0&maxtoreturn=20&codesystem=ICD9-CM

Parameters:
* matchvalue= a string for fulltext search or a query in Lucene syntax
* page= page number (starting from 0)
* maxtoreturn= number of elements per page
* codesystem= code system to query (mandatory)
* codesystemversion= code system version to query (optional)
* format= required format (e.g., "json")

2. Code System Version

Example: Entity *Immunoglobulina* in LOINC version 2.56:

- http://sti.iit.cnr.it/cts2framework/codesystem/LOINC/version/2.56/entities?matchvalue=immunoglobulina&page=0&maxtoreturn=20&format=json

3. Entity Description Read Service

Example: Read the detailed information of AIC code 19227038:

- http://sti.iit.cnr.it/cts2framework/codesystem/AIC/version/16.01.2017/entity/AIC:19227038

4. Association Query Service

Example 1: Existing cross-mapping associated to the ATC v. 2014 code "B02AA01".

- http://sti.iit.cnr.it/cts2framework/associations?list=true&codesystemversion=2014&sourceortargetentity=B02AA01&format=json

5. Entity Description Query Service

Example: List of LOINC codes (version 2.54) mapped to a local code system (e.g., Umbria Region):

- http://sti.iit.cnr.it/cts2framework/codesystem/LOINC/version/2.54/entities?page=0&maxtoreturn=250&matchvalue=LOCAL_CODE_LIST:Umbria&format=json

6. Export Service

Example: Export of AIC csv format, version January 2017:

- http://sti.iit.cnr.it/cts2framework/exporter?codesystem=AIC:16.01.2017&aictype=classe_h

To test the functionalities and suitability of STI, we recruited a sample of test users, belonging to some of the Italian Regions that already implemented the FSE infrastructure. On one hand, we provided special permissions to Domain Experts (e.g., General Practitioners and Laboratory technicians) in order to let them use both free functionalities (e.g., concept search, navigation of the resources, download) and the Cross-Mapping functionality, to create clinical/semantic mappings directly through STI. On the other hand, we asked regional technical referent users to query the Web service from their local application to make requests such as the ones provided above (e.g., to have the list of all the *map_to* codes in order to verify if some of their mappings changed the LOINC reference code). Figure 2 shows the cross-mapping performed by a Laboratory technician for the concept *Glucosio*.

VI. DISCUSSION AND CONCLUSIONS

This paper described the design and development of a bilingual (Italian – English) integrated terminology service, named STI, based on the CTS2 HL7 standard. The service includes for now the four code systems required by the FSE Law Decree, but it is open to the possibility to integrate further terminologies in the future.

Designing a terminology service is a non-trivial pursuit, especially when resources with different structures need to be integrated and available for different uses. This was the first issue of this work, as it required a personalized design and implementation for each code system uploaded into the STI KB. For example, LOINC has multiple informative axes, which were reported into both the main visualization screen (the six fundamental axes) and an openable window tagged with different labels. Nonetheless, importing LOINC into the service was challenging because its database structure changes as versions evolved. So, a preliminary normalization step was carried out to uniform names and values of the fields of the different versions. Moreover, dealing with AIC, as the system is released in four separate files, ETL procedures needed to be trained for importing each of them every time there is an update and checking if mappings to ATC are present in the new AIC files or if they need to be extracted by following the procedure described in Section IV.B.2). All the above mentioned issues are an obstacle to the flexibility and scalability of the service. Furthermore, it was not always easy finding updated versions of the four code systems, especially in computable format, such as csv files, and for some of them both master English and translated Italian files are not available (i.e., ATC). The chance to visualize ontology representation of the clinical terminologies is not usable for all the versions of the systems. This is an interesting possibility offered by the STI that needs to be improved in the future releases of the service. Efficiency and effectiveness of an EHR also depend on the possibility of unambiguously exchanging and understanding incoming information.

Semantic interoperability improves significantly thanks to the implementation of a terminology service, especially if it is compliant to a standard such as HL7 CTS2, which is widely adopted. The offered services (e.g., searching, querying, and cross mapping) are particularly useful when national or local code systems need to be linked to standard classification systems.

 ISBN: 978-1-61208-600-2

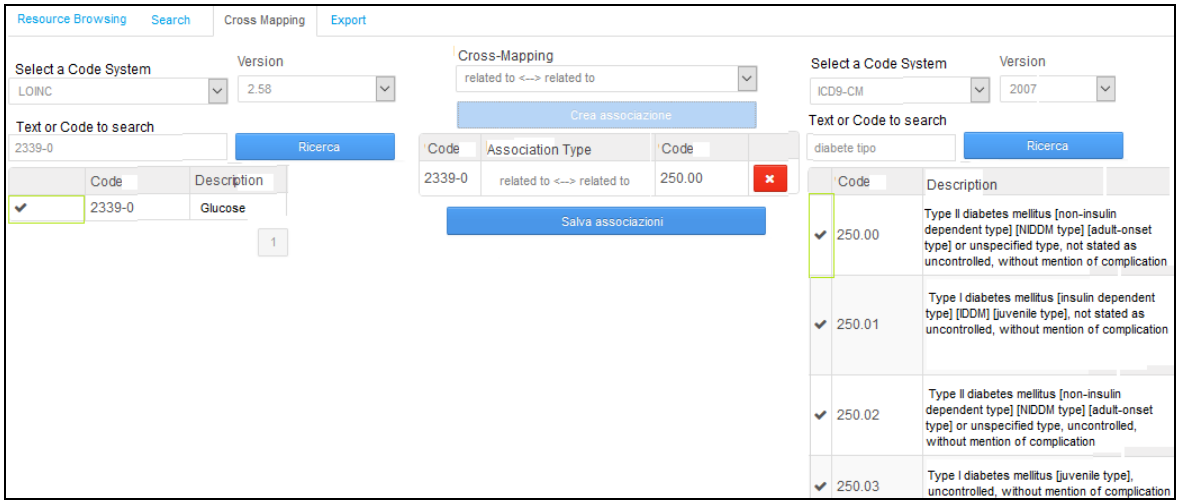

Figure 2. STI Web Application screenshots showing the Cross-mapping between the LOINC code 2339-0 “Glucosio” and ICD-9-CM.

This interoperability also strongly depends on the alignment between terminologies and their quality. This work shows the path that has been taken, also thanks to the recent advancements promoted by the law and by the AgID and CNR collaboration within the context of FSE projects, to align Italian FSE with international initiatives that promote the use of integrated management services of medical terminologies. Nonetheless, it is to be considered that the implementation of integrated terminology services is just the beginning of the process. In fact, the most important aspect in managing medical terminologies is the maintenance over time to update resources and coordinate processes such as transcoding, translation, and licensing. In fact, maintenance of such a system is the real challenge: systems change, errors are made, and the lifecycle of mappings and data must be considered. Sometimes, mappings can be contextual and absolute consistency is very hard to achieve. That evidences the need for a dedicated governmental authority to coordinate the entire process.

Among the several advantages provided by STI use in the Italian FSE frame there are:

- the possibility to share official terminologies, and their updates between the FSE central node and the services used by the local/regional FSE nodes;
- the possibility to configure policies (roles and authorizations) and to model the organization of the system (concerning production/editing of the terminological resources) through terminology management roles;
- the compliance of the data model and application services with the CTS2 standard (Normative Standard CTS2 Version 1.2);
- the delivering services of terminological resources with standard protocols and formats (json, CSV);
- the possibility to make advanced searches with personalized filters according to the code system selected, and to find additional semantic information by navigating their ontological graphs;
- the distribution as open source tool, with a GNU GPL license.

Some of these advantages and functionalities characterize STI if compared to existing CTS2 implementations, especially at national level. In fact, the terminology services cited in Section II.A, even if more sophisticated from a technical and architectural point of view (e.g., in the cited DiTAM service, the possibility to have many local terminology service nodes connected to the central DiTAM node in a federated network), are: proprietary, thus more difficult to be used by a Public Administration; less precise in the structuring and visualization of the code systems; and, to our knowledge, do not allow the access to the resources as Linked Data, or their ontological graphs; and, finally, they do not provide bilingual access as provided in STI.

Possible improvements that could be made on the STI in the future include: i) an extension of the service for including and managing also local code systems; ii) the definition of a general structure for importing and mapping in order to make the service more flexible. The ability to share, query and maintain official and up-to-date terminological artifacts using an accepted standard terminology service interface, such as STI will allow standard terminology content to be readily disseminated and validated, and becomes more useful as organizations (healthcare facilities, Regions, Ministry of Health, and national Standard Development Organizations) in the FSE context begin to undertake the enhancement and maintenance of terminologies to support language translations, jurisdictional extensions to standard code

 ISBN: 978-1-61208-600-2

systems, or maintenance and development of local terminologies, avoiding the proliferation of heterogeneous resources, and local tools and technologies to manage terminologies.

Finally, the creation of STI in the context of the Italian FSE, is not only a way to reach semantic interoperability, but it represents a better support to healthcare professionals for improving the quality of clinical data ensuring maximum benefits along the healthcare process and the cooperation among different healthcare providers.

ACKNOWLEDGMENT

This work is supported by the projects *Realizzazione di Servizi della Infrastruttura Nazionale per l'Interoperabilità per il Fascicolo Sanitario Elettronico* and *Realizzazione dei servizi e strumenti a favore delle Pubbliche Amministrazioni per l'attuazione del Fascicolo Sanitario Elettronico* funded by the Agency for Digital Italy. A special thanks goes to LINK Management and Technology for the support in the development of STI.

REFERENCES

[1] M. Ciampi, A. Esposito, R. Guarasci, and G. De Pietro, "Towards Interoperability of EHR Systems: The Case of Italy" ICT4AgeingWell, 2016, pp. 133-138.

[2] E. Cardillo, M. T. Chiaravalloti, and E. Pasceri, "Healthcare Terminology Management and Integration in Italy: Where we are and What we need for Semantic Interoperability" European Journal of Biomedical Informatics, vol. 12 (1): pp.en84-en89, 2016.

[3] Prime Ministerial Decree [Law (general)] n. 178, 29 September 2015, "Regolamento in materia di fascicolo sanitario elettronico. (15G00192)" GU Serie Generale n.263, 11-11-2015.URL: http://www.gazzettaufficiale.it/eli/id/2015/11/11/15G00192/sg [accessed: 09-15-2017].

[4] ANSI/HL7 V3 CTS R2-2015, HL7 Common Terminology Services - Service Functional Model Specification, Release 2, February 2015, URL: https://hssp.wikispaces.com/specs-cts2 [accessed: 09-18-2017].

[5] European Community, "Directive 2011/24/EU of the European Parliament and of the Council of 9th March 2011 on the application of patients' rights in cross-border healthcare" [Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:088:0045:0065:EN:PDF].

[6] Semantic Interoperability for Health Network FP7-ICT-2011-7, URL: http://www.semantichealthnet.eu/ [accessed: 09-15-2017].

[7] C. Daniel, D. Ouagne, E. Sadou, K. Forsberg, M. Mc Gilchrist, E. Zapletal, N. Paris, S. Hussain, M. Jaulent, D. Kalra, "Cross border semantic interoperability for clinical research: the EHR4CR semantic resources and services," AMIA Jt Summits Transl Sci Proc. 2016; 2016, pp.51–59.

[8] Trillium Bridge II - Reinforcing the Bridges and Scaling up EU/US Cooperation on Patient Summary, H2020-EU.3.1.5. - Methods and data, URL: http://www.trilliumbridge.eu/ [accessed: 09-18-2017].

[9] CTS2 Development Framework, Mayo Clinic Informatics, URL: https://github.com/cts2/cts2-framework [accessed: 09-18-2017].

[10] Fast Healthcare Interoperability Resources (FHIR), Health Level 7 Specification, URL: https://www.hl7.org/fhir/ [accessed: 09-18-2017].

[11] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology" Nucleic Acids Res: vol. 1; 32(Database issue); pp. D267–D270, 2004 DOI: 10.1093/nar/gkh061.

[12] J. Grosjean, G. Kerdelhué, T. Merabti, and S. J. Darmoni, "The HeTOP: indexing Health resources in a multi-terminology/ontology and cross-lingual world," 13th EAHIL Conference, 2012, ceur-workshop, vol.952, 2012. URL: http://ceur-ws.org/Vol-952/paper_17.pdf [accessed: 09-15-2017].

[13] J. Pathak, H. R. Solbrig, J. D. Buntrock, T. M. Johnson, and C. G. Chute, "LexGrid: A Framework for Representing, Storing, and Querying Biomedical Terminologies from Simple to Sublime" Journal of the American Medical Informatics Association: JAMIA. vol. 16(3), pp.305–315, 2009, DOI:10.1197/jamia.M3006.

[14] D. K. Sharma, H. R. Solbrig, E. Prud'hommeaux, K. Lee, J. Pathak, and G. Jiang, "D2Refine: A Platform for Clinical Research Study Data Element Harmonization and Standardization," AMIA Jt Summits Transl Sci Proc. 2017; 2017, pp. 259–267.

[15] L. Zhou, H. Goldberg, D. Pabbathi, A. Wright, D.S. Goldman, C. Van Putten, A. Barley, and R. A. Rocha, "Terminology modeling for an enterprise laboratory orders catalog", AMIA Annu Symp Proc., 2009, pp.735-739.

[16] K. J. Peterson, G. Jiang, S.M. Brue, and H. Liu, "Leveraging Terminology Services for Extract-Transform-Load Processes: A User-Centered Approach," AMIA Annu Symp Proc. 2017 Feb 10;2016, pp.1010-1019.

[17] Standard Terminology Services (STS), PHAST, URL: http://www.phast-association.fr/accueil_doc_sts/ [accessed: 09-15-2017].

[18] C. Seerainer and S. W. Sabutsch, "eHealth Terminology Management in Austria" Studies in health technology and informatics., vol.228, pp-426-30, 2016.

[19] Terminology Server, University of Applied Sciences and Arts, Dortmund, URL: http://www.wiki.mi.fh-dortmund.de/cts2/index.php?title=Hauptseite [accessed: 09-15-2017].

[20] DiTAM, Codices S.r.l.. 2014. URL: http://www.codices.com/prodotti/ditam.html. [accessed: 09-18-2017].

[21] S. Canepa, S. Roggerone, V. Pupella, R. Gazzarata, and M. Giacomini, "A Semantically Enriched Architecture for an Italian Laboratory Terminology System," IFMBE Proceedings, XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013. Berlin: Springer; vol. 41, 2013, pp. 1314–1317.

[22] Liferay 6.2., URL: https://www.liferay.com/downloads . [accessed: 03-11-2017].

[23] Pentaho Data Integration (ETL) a.k.a Kettle, URL: https://github.com/pentaho/pentaho-kettle . [accessed: 03-11-2017].

[24] Virtuoso, OpenLink Software, URL: https://virtuoso.openlinksw.com/ . [accessed: 03-11-2017]

[25] CKAN, URL: https://ckan.org/ . [accessed: 03-11-2017].

[26] Autorizzazione all'Immissione in Commercio (AIC), Agenzia Italiana del Farmaco. URL: http://www.agenziafarmaco.gov.it/en . [accessed: 09-18-2017].

[27] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications". Nucleic Acids Res. 2011 Jul;39, pp.W541-5. Epub 2011 Jun 14. URL: https://bioportal.bioontology.org/ [accessed: 09-15-2017].

 ISBN: 978-1-61208-600-2

semantify.it, a Platform for Creation, Publication and Distribution of Semantic Annotations

Elias Kärle, Umutcan Şimşek and Dieter Fensel

STI Innsbruck, University of Innsbruck,
Technikerstrasse 21a, 6020 Innsbruck, Austria
{elias.kaerle, umutcan.simsek, dieter.fensel}@sti2.at

***Abstract*—The application of semantic technologies to content on the Web is, in many regards, important and urgent. Search engines, chatbots, intelligent personal assistants and other technologies increasingly rely on content published as semantic structured data. Yet, the process of creating this kind of data is still complicated and widely unknown. The *semantify.it* platform implements an approach to solve three of the most challenging question regarding the publication of structured semantic data, namely: a) what vocabulary to use, b) how to create annotation files and c) how to publish or integrate annotations within a website without programming. This paper presents the idea and the development of the *semantify.it* platform. It demonstrates that the creation process of semantically annotated data does not have to be hard, shows use cases and pilot users of the created software and presents where the development of this platform or alike projects lead to in the future.**

Keywords–schema.org; semantic annotations; Semantic Web; annotation platform; software as a service.

I. INTRODUCTION

The creation of annotations for Web content should be neither complicated nor painful, but intuitive and easy for all content creators or Web page editors. Not too long ago, the challenge was to have a well structured and beautiful looking website. This was solved by the establishment of content management systems (CMS). Now, as the focus on the Web shifts away from content- and design based websites towards well structured, high quality content [1], [2] the demand for a CMS like tool to create such structured content grows.

The high demand for annotated data originates in the development of a layer on top of the Web as we know it, called the *headless Web* [3]. Within this layer, the number one consumer of content is no longer a human browsing the Web, but machines. These machines browse the Web with much higher velocity and accuracy and aim to take over search efforts for humans. Intelligent personal assistants (IPA), like Amazon's Echo [4], Apple's Siri [5], Google's Allo [6] or Microsoft's Cortana [7], answer questions, asked by humans, based on high quality structured information from the Web. Chatbots, too, aim at replacing humans as Q&A (question and answer) counterparts by retrieving answers from high quality data on the Web. The change in the user interface of popular search engines shows that they also try to answer users' demands directly within the search engine website, without the need to lead the user to different, linked, pages. See Figure 1, for an example of a search result displayed inside a search engine website.

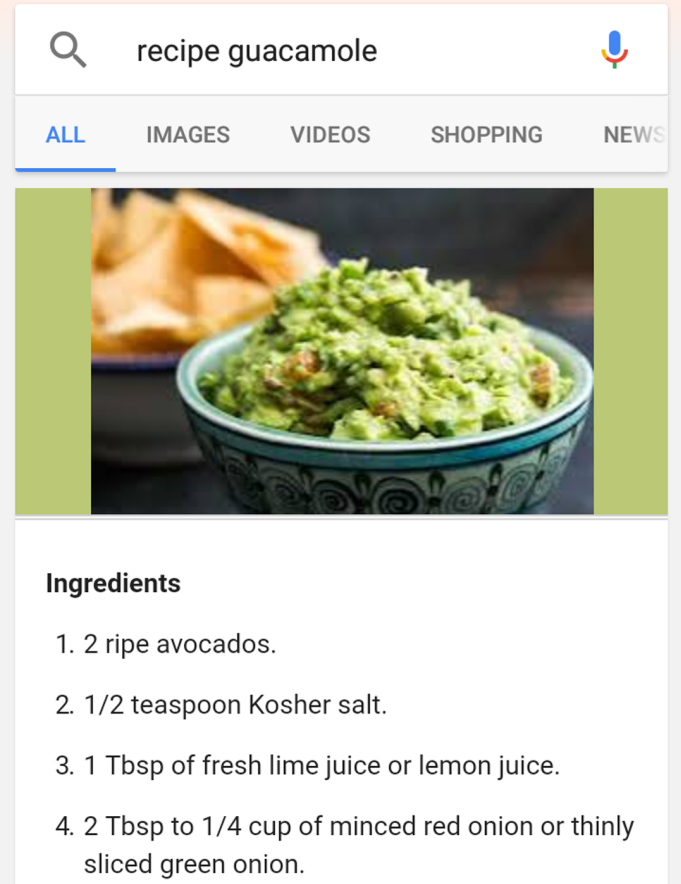

Figure 1. Example of a search for "guacamole recipe" with the result displayed inside the search engine website of Google.

To structure the content on the Web, there is a variety of vocabularies to choose fro,m but the most widely acknowledged one [8], [9] has proven to be schema.org [10]. Schema.org is an initiative launched by the "big 4" search engine providers Bing, Google, Yahoo! and Yandex in 2011. It is a collection of terms to describe *things* on the Web in a structured way. It is embedded into the HTML [11] source with either RDFa [12], Microdata [13] or JSON-LD [14], [15]. In this work, we will only focus on the latter one. An analysis conducted by Kärle et al. [16] has shown that schema.org is widely distributed, but used mostly in an incomplete or wrong way. But why is the creation of annotations so hard in the first place? The root of

 ISBN: 978-1-61208-600-2

the problem can be summarized in three questions: (1) which vocabulary to choose, (2) how to create JSON-LD files and (3) how to publish JSON-LD files.

With *semantify.it* [17] we provide a Web application whose main purpose is to make the creation of annotations easy and intuitive. But it is not only a platform for creation and storage of annotations but also to validate, edit, analyze and publish annotated data.

The Web based software is free of charge and anyone can register, start creating annotations based on pre-built forms made by domain experts or simply upload annotations. The generator is easy and intuitive to use and the resulting JSON-LD files are stored on the server. From there, they can be fetched and integrated into existing websites with the help of content management system plugins or by a unique URL call. In addition to the static data, the platform contains an extension framework, through which applications that map external data sources to schema.org, can push dynamically created annotations to the *semantify.it* platform.

The remainder of this paper is structured as follows: Section II lists work related to the approach presented in this paper and states the motivation to build the *semantify.it* platform. Section III shows the technical approach and Section IV describes implementation details. Section V presents the results of the work on the platform and Section VI concludes the paper and gives an outlook to future work and additional projects.

II. Related Work & Motivation

In this Section, we review the existing annotation tools and frameworks as well as CMS extensions and explain our motivation for developing the *semantify.it* platform to facilitate the annotations process. Annotation of unstructured content on the Web has drawn a lot of interest from the semantic Web community. Since schema.org emerged in 2011, all parties on the Web have gained major motivation for annotating their content, especially for the benefits coming from the support of the major search engines to structured data markup. The recent developments in the intelligent personal assistants (IPA) and chatbots also increased the importance of semantically described structured data on the Web. The content on a webpage can be semantically enriched by embedding the annotation of the content to HTML source in formats such as JSON-LD, Microdata and RDFa. However, without proper tool support, the structured content publishing process can be very challenging for the end-user.

There are many annotation tools and frameworks in the literature with different levels of automation (e.g., automated annotation with natural language processing) [18]. Comprehensive surveys of such tools and frameworks can be found in [19], [20] and [21]. These annotation and knowledge extraction tools aim to semantically enrich documents and to enable semantic search and reasoning. However, these tools did not find major practical use for annotation of webpages, since they do not create full annotations, but mostly recognize and link entities in text. The technical challenge of embedding annotations into the webpage has been tackled by extensions/plugins for popular CMS [22] [23]. Our approach decouples the generation and publication of the annotations, which allows experts who do not necessarily have access to the administration panel of the CMS to create annotations. Then, our generic CMS extensions can find and inject annotations to webpages. Since the CMS extensions share a common PHP (PHP is a server-side scripting language) API (application programming interface) for communicating with our platform, the CMS specific development effort is kept minimal. Besides the creation and publication, another major challenge of the annotation process remains mostly untouched. Schema.org is a relatively large vocabulary with many types and properties and it is not easy for an end-user to pick relevant types and properties for annotations in a certain domain. Morover, CMS extensions generally support a predefined set of types and properties (Mostly Article and BlogArticle with mappings from metadata fields of CMS posts to corresponding properties of the aforementioned types). An exception could be RDFaCE [22], which allows users to pick desired types from the entire schema.org vocabulary, but the selection is only limited to types, while the properties and ranges cannot be restricted. Additionally, with our approach, we enable the creation of annotations based on the frequently changing data, which is not feasible to annotate manually with a CMS extension. The mappings from an APIs data structure or a relational database schema to the schema.org vocabulary should be done. This task requires major development on the CMS side.

We propose the *semantify.it* platform which facilitates creation, validation and publication of structured data on the Web. The annotations can be done manually via an editor that is generated automatically based on a domain specification (a specific subset of the schema.org vocabulary (see [24] for details)) or automatically through an extension that maps the data structure of external data sources to a domain specification. The data from the mapped data sources then can be pushed to the system via an open RESTful (REST stands for representational state transfer) API. Creating annotations against a domain specification (e.g., Google structured data guidelines) helps end-users to ensure that their annotations are compliant with search engines' structured markup guidelines. We are also implementing a rule based validator for semantic validation of the annotations. The publication of the annotations are done by simple generic extensions that we develop for popular CMS, which merely maps generated annotations to Web pages. Additionally, our open RESTful API allows application developers to reuse the annotations hosted in *semantify.it*, without crawling.

A recent effort from the W3C (World Wide Web Consortium) Web Annotation Working Group, the Web Annotation Data Model [25] and Vocabulary [26], aims to standardize the annotations on the Web. The ultimate goal of the standard is to open and decentralize the comments on the Web content. It also allows more fine-grained annotations, meaning that it is possible to make comments on a part of the content. This effort does not relate to the purpose of our platform directly, since it is actually a vocabulary for describing the annotations. Nevertheless, the idea of separating annotations from content and publishing them on-demand is somewhat parallel to our vision.

To the best of our knowledge, there is no such platform that generates, validates and publishes annotations in a holistic way. By decoupling the annotation creation and publication, we enable content creators who do not have extensive knowledge about schema.org to benefit from semantic annotations, since they can be externally generated by experts and be stored on

 ISBN: 978-1-61208-600-2

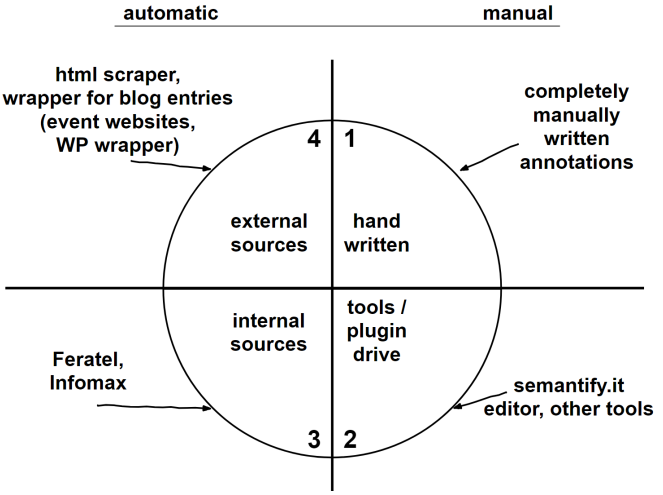

Figure 2. The four different types of content creation. Every quarter stands for a different type how an annotation can be created.

semantify.it platform.

III. METHODOLOGY

In Section I, we introduced the three major challenges when it comes to the annotation of content on the Web. The first and probably most important question to answer is, which vocabulary to use. Due to its growing importance and distribution, we choose to support the *schema.org* vocabulary [9]. But still, there are hundreds of classes and properties to choose from, which makes it very hard for inexperienced users to select the right set of classes and properties to annotate certain Web content. To solve that challenge, *semantify.it* provides a set of recommendation files for different domains or use cases which define, case specific, which classes and properties are recommended or even required to create proper annotations. Since those recommendation files always target a certain domain they are called *domain specification* or *DS* in short [24]. The second question is how to create proper annotation files in the recommended JSON-LD format [27] (JSON-LD is a JSON-based data format for linked data purposes, which became a W3C recommendation in 2013 [28]). To answer that, *semantify.it* provides the user with specific editors for specific domains or use cases. The editors are based on the selected classes and properties from question one and are generated dynamically every time the user creates an annotation. The third question *semantify.it* addresses is the publication of annotations. Of course, the annotation file can be copy-pasted into a script tag of a website, but most Web- or CMS users are not able to fulfill this task. So *semantify.it* stores all created annotations and provides them through an API and also offers a number of plugins to popular content management systems to automatically retrieve the annotation files from the *semantify.it* server and inject them into the website.

A. Creation

As depicted in Figure 2, we distinguish between manual and automatic annotation creation. The two options have two more distinctions. Manual annotations can either be completely handwritten, with a text editor (first quarter), or tool driven, like with the *semantify.it* editor (second quarter). Automatic annotations can be divided, from a service point of view, into internal sources and external sources. We talk about internal sources if the structure of the data is known or agreed upon and the structure is maybe even protected by a service level agreement (third quarter). External sources are such, where the structure is unknown and has to be determined based on an HTML source. The structure can change any time and no agreement between the data provider and the annotation creator exists (fourth quarter).

1) Manual annotation creation: The manual annotation creation process is, as mentioned above, driven by editors based on domain specifications. The concept of defining subsets of schema.org annotations for domain specific usage was first presented by Şimşek et al. in [24] and was adapted for the usage in *semantify.it*. To build the DS files, the platform features an editor, which lets the domain expert select classes, properties and sub classes just by clicking. No source code has to be written at all. The DS files are then saved and accessible to all users over the annotation editor interface. When starting a new manual annotation, the user selects a DS on which the annotation will be based and gets presented with the corresponding editor. The editor for annotation creation looks like an ordinary HTML form and hence gives the user a good and familiar usability. If all required fields are filled the user can proceed by clicking the ”generate” button and gets presented with the annotation source code in JSON-LD format. This source code can then either be copy-pasted or stored on the *semantify.it* platform for further usage.

Manual creation of annotations with an editor is only one way to use *semantify.it*. Of course the platform does not cover domain specification for all use cases. So someone might create annotations in a different way but still wants to utilize *semantify.it* for storing and distributing those annotations. For this case there is an upload functionality where one or multiple annotations can be posted to the platform where they then are treated exactly as the annotations created with an editor.

To introduce further ways of deploying annotations to the *semantify.it* platform, we first have to define a distinction between three different types of content, which are *static content*, *dynamic content* and *active content*. Static content hardly changes after having been produced once. For example on a hotel website it is mainly the hotel’s core data: name, address, phone number, email and alike. Dynamic data changes frequently or even constantly. To stick with the hotel example: room availabilities, prices or specialties on the daily menu count as dynamic data. Active data is information about interfaces to interaction software on a website, like, for example, an internet booking engine’s API. The manual creation of annotation with the *semantify.it* editor targets mostly static data. Due to its nature it makes hardly any sense to annotate dynamic data manually.

2) Automatic annotation creation: For the use with dynamic data, *semantify.it* offers, similar to the upload functionality, the functionality to send annotation files to an API to be stored on the platform. So annotations created automatically, for example by a wrapper software, can be stored on and distributed by *semantify.it* as well. Some of those wrappers are even integrated into the platform and provided to the user as *extensions* (see III-D). To make use of such a wrapper extension, the user has to activate the extension, provide it with its credential to the given data source and set the frequency

 ISBN: 978-1-61208-600-2

for the wrapping process. Then the data of the source, a WordPress blog or a destination management software, is crawled, mapped and stored on *semantify.it* recurrently.

B. Validation

Another part of annotation creation is the annotation validation. *Semantify.it* offers a validation feature based on the ideas mentioned in [24] to give the user feedback if the information he was entering makes sense. The validation process performs a semantic validation where, based on validation rules, a check is performed if the data entered makes sense according to the rules defined by a domain expert. So for every DS mentioned above, there is the option to create a corresponding domain rule file, or DR for short, to perform semantic validation. Currently the prototypical validation feature is being improved on the platform. The developments on the rule editor are ongoing.

C. Storage & Publication

As mentioned above, most content creators are not able to copy-paste annotation files into their content management systems. So *semantify.it* provides an infrastructure for storage, maintenance, analysis and publication of annotations. Every file created, uploaded or stored through the public API, is assigned to a concept called *website*. A *website* is associated with a *user* who can create several of those *websites*. Every *website* has an API key, which is used to fetch annotations from or store annotations to said *website*. Mostly the *website* concept of *semantify.it*, as a collection of annotations, correlates with a real website where the annotations belong to that is the reason for the naming. An annotation is uniquely identified by a nine alphanumeric character long URL safe hash code. To retrieve the annotation from the *semantify.it* server the user just has to enter the shortener URL [17] and append the hash code. The response is a plain JSON file containing the corresponding annotation in JSON-LD format. On the dashboard the platform shows all annotations grouped by *website*. Every annotation has the possibility to be previewed or edited. Editing works by loading the corresponding editor, populating the form with the content from the annotation and overwriting the old annotation when the user is done editing. *Semantify.it* provides an analytics feature for (so far) basic statistics about the number of classes and properties annotated and the overall number of facts stored for each *website*. This functionality will be extended in the future (see Section VI). There are several ways to publish annotations stored on *semantify.it*. For static data it might make sense to fetch every annotation separately for a webpage by the hash identifier. For dynamic data there are two possibilities: (1) if an annotation gets stored on *semantify.it*, the software checks for a valid schema.org/url property. If that property exists it gets URL-encoded and stored as a retrieval key for that annotation file. The annotation can then be fetched by calling the shortener URL followed by "url/" and the URL-encoded content of the schema.org/url parameter. This method makes sense when a Web master decides to automatically annotate a huge number of blog entries and store them on *semantify.it*. The annotations can then be retrieved with a CMS plugin (see IV-C) where each annotation file is identified by its encoded URL. (2) with a custom identifier, called CID. The API call to send annotations to *semantify.it* offers the possibility to add an optional CID parameter for each annotation. Annotations stored with a CID can be fetched from the server by calling the shortener URL followed by "cid/" and the value of the custom identifier. This makes sense for systems where Web content is stored in a database and then injected into a Web page based on a CID. Those Web pages can be annotated automatically and the annotations can be injected, just as the Web content itself, by the corresponding CID. As part of the publication functionality, *semantify.it* provides a number of plugins for popular content management systems (so far for WordPress [29] and Typo3 [30], but Joomla [31], Drupal [32] and more are in the pipeline). Those plugins can be downloaded from the CMS provider's plugin repository. Therefore, the *website's* API key has to be stored in the plugin's settings and the configuration has to be set to either load the annotations manually per Web page or automatically by Web page URL. Then, on every Web page call, the annotations get fetched from *semantify.it* and injected into the Web pages created by the CMS.

D. Extensions

Besides the possibility to create annotations manually and to use the service as a storage, maintenance, analytics and publication platform, *semantify.it* also offers an extension functionality, which targets automatic annotation creation. An extension is actually a piece of standalone software, which is integrated into the *semantify.it* platform. A user can optionally activate extensions and configure them individually. Extensions are developed by the *semantify.it* team or can be suggested by external developers through Bitbucket [33] or Github [34]. Some examples for extensions are listed below.

1) Data mapping: a lot of websites obtain their content from external sources, which have public APIs. For example a destination management organization's (DMO) website contains data about room offers or hiking paths and the data is provided by different vendors through their APIs. If the DMO wants the content to be annotated it either has to convince the data provider to annotate all the data (which is probably hard) or use the corresponding *semantify.it* extension. The extension requires the API access data of the user and then starts to crawl the data, map it to schema.org and store the annotations on *semantify.it*. A simple plugin can then pull the annotation from *semantify.it* and inject the right annotation to the corresponding Web page. An example for the use of data mappings for massive annotations of destination management organizations' websites can be found in [35].

2) WordPress article annotation: another example for an extension is the annotation of blog articles in WordPress. Currently there are no plugins which annotate pages or blog entries in WordPress directly. So, if an author decides to create annotations for all his old articles this can become very painful. So *semantify.it* provides an extension, which crawls all blog articles of a given website, maps the relevant content to schema.org and stores the annotation file on *semantify.it*. A plugin, like the one explained above, can then fetch the annotation and inject it into the corresponding WordPress Web page. The same could work for other blog systems too.

IV. Implementation

Semantify.it was designed and implemented to be delivered as a software as a service or SaaS [36]. To support version control during the development we make use of the free

 ISBN: 978-1-61208-600-2

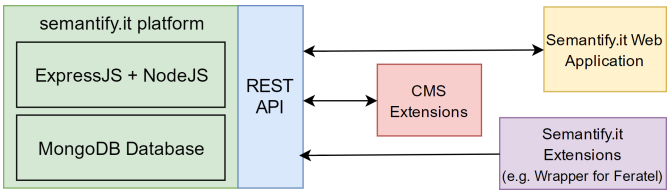

Figure 3. Modules of the *semantify.it* platform.

and open source version control software Git [37], hosted on the platform Bitbucket. Through short commit cycles, sophisticated branching and meaningful commit messaging the code is kept as manageable as possible, relatively easy to maintain and easy to roll back in case grave errors should be detected only after a release.

The reference architecture of the *semantify.it* is depicted in Figure 3. In the following subsections, we will explain the main modules in our architecture and the communication between external applications and the *semantify.it* platform.

A. Platform Core

The platform core has been implemented with NodeJS and the ExpressJS Web application framework, which allowed us to create a lightweight Web based platform with a RESTful API. We adopt the document-oriented database system MongoDB for persistance. A document-oriented database is a natural decision when working with JSON-LD files, since they can be stored directly as documents. This way, we can serve the annotations as they should appear in the HTML source of a Web page. Another possibility such as using a triple store would be more suitable for developing semantic Web applications based on the stored annotations, however this would make querying the annotations to obtain JSON-LD documents to be embedded into a Web page more challenging. Even though it would be possible to query a single blank node and all the other nodes that are connected to this blank node with the help of property paths, this would still be tricky since an annotation is typically a RDF graph, which consists of many blank nodes connected to each other. Therefore, referring to a specific node without preprocessing the annotations would not be possible. This gets even more complicated when an annotation contains two disconnected graphs. In this case storing each annotation in their own named graph may be a solution. Nevertheless, the main purpose of the *semantify.it* platform is to serve the annotations to Web pages, therefore we handle the annotations as JSON-LD documents, rather than graphs. The hosted annotations can be then retrieved via the REST API and stored in a triple store in a desired way.

The platform manages the annotations in relation with websites, organization and users. We define the concept of website in our data model, which can host multiple annotations. Every website has a unique API key. An organization can have multiple users and they can manage multiple websites that belong to their organizations. External applications can perform CRUD operations on the core platform via the RESTful API. The API key of a website is used by external applications for creating and retrieving annotations. More sensitive operations, such as updating and deleting annotations require additional security measures. In this case, the authentication of users is handled by JSON Web Tokens (JWT).

B. Web Application

The Web application is developed as an interface to the core platform. In the frontend, the application uses HTML5, CSS, Javascript and Material Design [38] elements. It communicates with the RESTful API of the platform core with JQuery. The application currently supports the fundamental functionality such as user registration, website management, annotation creation based on domain specific editors, domain specification and bulk upload of annotations. Additionally, users can see certain statistics about their websites (e.g., number of annotations, number of statements, number of annotation requests). The Web application has access to all the routes defined in the RESTful API.

C. CMS Plugins

We develop two plugins for the popular CMSs Wordpress and TYPO3. According to Web Technology Surveys [39], Wordpress is the most widespread CMS worldwide, and TYPO3 is very common in German speaking countries. Therefore, plugins for these CMS in the initial phase covers many use cases. Both extensions use a common PHP API to communicate with the RESTful API of *semantify.it* platform. The front-end development of the plugins is tailored for each CMS since they vary in plugin architecture. The only configuration the CMS plugins need is the API key of a website on *semantify.it*. The plugins have two main functionalities; (a) they allow the content creator to tie a specific page/post with a specific annotation hosted in the *semantify.it* platform and (b) the plugin can use the URL of a page and retrieve an annotation that has the same URL as the value of the *url* property. This feature is very useful in most cases, however, a user can always opt out from using it and select annotations manually.

D. Extensions

One of the most important challenges of semantic annotations on the Web is their maintenance. In many cases, the important data usually changes frequently, therefore keeping the annotations up to date is an important task. For instance in the tourism domain, accommodation offers can change on a daily basis. In the *semantify.it* platform, we offer an extension mechanism where the data from external data sources (e.g., Feratel) are mapped to domain specifications and annotations are generated automatically with a specified frequency. Annotation of frequently changing data through wrappers has been described in [35]. We also create wrappers as extensions to the *semantify.it* platform, which can be activated by users when needed. The mapping is currently done within the wrapper's business logic, but we plan to adopt an RDF Mapping Language (RML) [40] based approach in order to increase re-usability (see Section VI). The automatically generated annotations are stored in the database with a unique custom ID (CID). This ID is generated based on the external data source's entity identification scheme. For instance, Feratel uses UUID [41] for identifying entities such as accommodations. Based on this, we create CIDs in "FeratelID-languageCode" format, since we want to identify annotations in different languages separately. The annotations then can be matched on the client's side with the corresponding webpage where the content about an entity resides in a certain language.

The main challenge of the extension development lies in the mapping of custom data formats and structures to schema.org

 ISBN: 978-1-61208-600-2

vocabulary. In some cases the entity types in the external source's data model may be too specific for schema.org vocabulary. We overcome this challenge by using a suitable more generic type from schema.org. Another challenge is that some information may be given in an unstructured way in the external data source, which makes it tricky to map and extract programmatically. In such cases, we try to find patterns in the content and write suitable extractors. If this is not possible, we simply ignore that content.

V. Results & Use Cases

This section will show the use cases of our implementation and present statistics obtained from the initial usage.

The platform started with one ski school as a pilot. Meanwhile the three destination marketing organizations (DMO) of Mayrhofen [42], Seefeld [43] and Fügen [44] are testing *semantify.it* as pilots and the umbrella organization of Tirol's tourism organization, Tirolwerbung [45], is about to use *semantify.it* with a wrapper extension. Besides that, several private websites are working with *semantify.it* and providing feedback. A more detailed description of those use cases will be presented in V-B. Also, the WordPress [46] and the Typo3 [47] plugin are used already by pilots and deliver thousands of annotations every day. Section V-A gives more details about that.

A. Results

At the time of the evaluation, *semantify.it* was hosting 31 users in 27 organizations maintaining 42 websites. There were 37.597 annotation files stored, containing more than three million annotation statements (triples), which where requested over the API more than 82.000 times in the time span between April 5th, 2017 and June 14th, 2017. For the better understanding of the annotation file to annotation request ratio, it is important to mention that not all pilots test the whole work flow of *semantify.it*. Some are testing the bulk upload feature through data wrapper extensions, which leads to a huge number of annotation files, but the CMS extension not yet, which explains the relatively low annotation request number. Others created their annotation files manually with the *semantify.it* editor or a text editor but use the CMS extension, which leads to only small number of annotation files but a big number of annotation requests. Every page call on the CMS extension user's website triggers one annotation request on the *semantify.it* platform. So only the pilots having installed the CMS extension contribute to that number. As soon as all pilots use any form of CMS extension, the number will increase drastically. Currently the Typo3 plugin counts 127 downloads (which are not unique per website) and the Wordpress plugin counts less then 10 active installs.

To provide SSL (Secure Sockets Layer) capabilities to the users, *semantify.it* traffic is channeled through Cloudflare [48], a content delivery network. A picture of Cloudflare's analytics service shows the accesses to *semantify.it* (over UI as well as over the API) in the time from April 22nd to May 20th (see Figure 4).

To find out if loading annotations for websites from the *semantify.it* platform is acceptable in terms of response time we performed a series of response time measurements over a testing website [49]. The average response time was at around 150ms, which is an acceptable loading time for external scripts.

B. Use Cases

To test the functionality and the operational readiness we applied several different use cases to *semantify.it* and tried to find pilots for all four annotation creation scenarios described in Figure 2. We created and uploaded annotations manually and distributed them via the CMS extensions and we used annotations, which were created automatically and were uploaded to *semantify.it* vie the API. Those scenarios will be described below.

1) Manually created annotations: The first pilot of *semantify.it* was a ski school from Switzerland. Their website consists of 64 sub pages of static, rarely changing, content. For the purpose of being a *semantify.it* pilot, all annotations were created manually and uploaded through the platforms upload-feature. The total count of annotation statements in all the 64 annotation files is 5312, which means that there are 5312 facts or triples stored on *semantify.it*. The website uses a Typo3 CMS and has the *semantify.it* plugin installed. The administrators went through all the 64 sub pages and selected the corresponding annotations manually. This use case matches scenario one in Figure 2.

A use case for scenario two from Figure 2 was a hotel pilot. The annotations for the hotel, the included restaurant and some events were made with the *semantify.it* editor and integrated into the hotel's website with the Wordpress plugin.

2) Automatic annotation creation: A use case for automatic annotation creation (Figure 2, third quarter) was the mapping of Feratel's [50] tourism destination data into schema.org (as described in [35]). The thousands of annotation files of the DMOs of Mayrhofen [42], Seefeld [43] and Fügen [44], first stored in a file system, made a perfect use case for *semantify.it*. So we extended the existing wrapper and now every night all the data for the corresponding website from the Feratel system, annotated with schema.org, is uploaded to *semantify.it*. For the three DMOs mentioned above there are currently around 22,000 annotation files containing 3.9 million annotation statements. The annotation files, identified by a UUID stored as CID, are replaced if they already exist, otherwise newly created. The CMS plugin, which is not made by us but by the DMOs' Web agencies, is not ready yet. But we could find out that *semantify.it* can easily cope with thousands of annotation files and millions of annotation statements and the performance of the upload API scales.

A similar use case is the example of Tirolwerbung. Their Web agency maintains a self made CMS with an API to the database. We built wrappers for various different domains (hiking routs, ski resorts, accommodations and others) and now daily crawl the API to then store the resulting annotation files (around 6,000) on *semantify.it*. As in the previous example the annotation retrieval software for the CMS is not yet finished. The annotation files are identified by a CID with which they are also going to be fetched by the CMS plugin.

Another use case is the annotation of a corporate blog with around 220 entries. As an example for scenario four in Figure 2 we wrote a script, which scraped the content, mapped it to schema.org and stored it on *semantify.it*, which led to 14,191 annotations statements in 223 annotations files. The blog is built on Wordpress and through the use of the plugin the annotations are injected into the blog's HTML. In this use case the automatic annotation retrieval by URL property (described

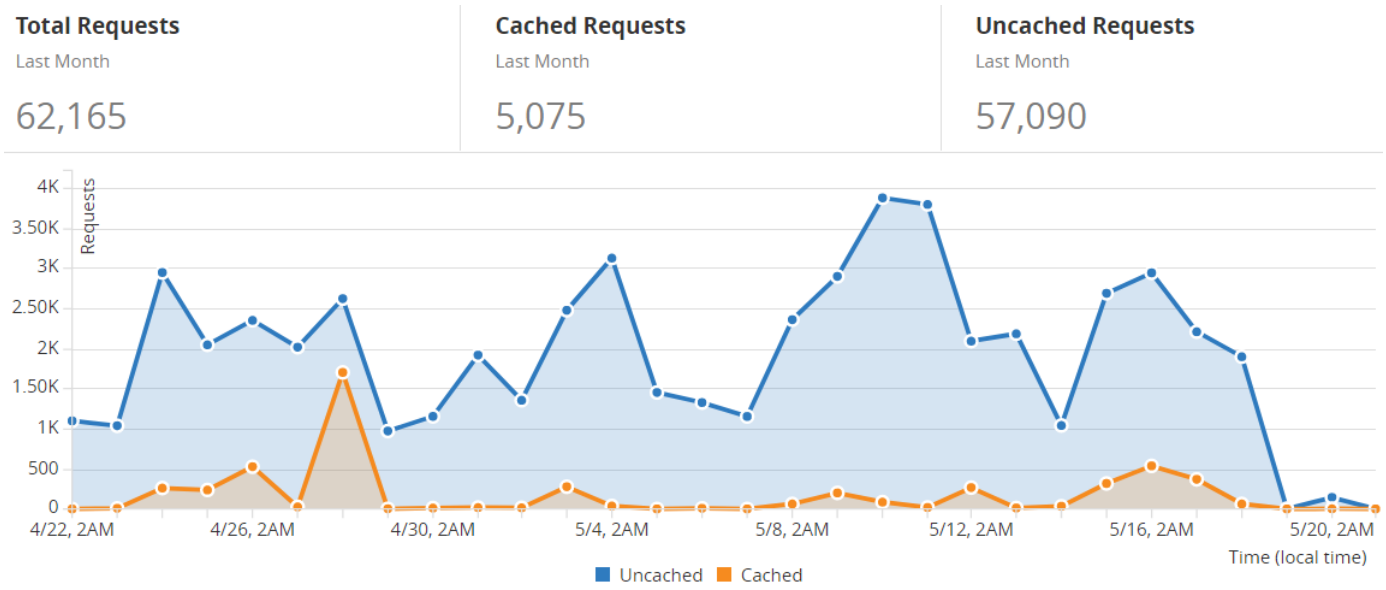

Figure 4. Access statistics of *semantify.it* for the one month time span from April 22nd to May 20th.

in III-C) comes into play. Thus, every annotation is integrated into the blog automatically and the administrator does not have to spend time assigning annotation files to Web sites.

VI. Conclusion & Future Work

This section concludes the work on the *semantify.it* platform, wraps up the outcome and gives an outlook into the future of developments on the software.

A. Conclusion

This paper describes the *semantify.it* platform, a multi purpose software-as-a-service to create, store, validate, publish and analyze semantic data. The easy-to-use interface and the comprehensive API make it easy to generate and store semantic annotations. The extension system, which generates annotated data out of different data sources makes annotations even more accessible for different users and purposes. Plugins to popular content management systems make the usage and publication of this structured data simple for non experienced users. Even though the individual parts of *semantify.it* might not be complete novelties, the idea of a holistic platform for creation, publication and distribution of semantic annotations is novel. As a proof-of-concept, use cases from the tourism field in Austria, Germany and Switzerland show that the *semantify.it* platform is capable of handling real life traffic reliably.

B. Future Work

There is still a lot to be done to make the creation and publication of semantic annotations easy and intuitive. Our efforts in enhancing *semantify.it* go in several directions, which will be shortly described below. To ensure the usability of the *semantify.it* user interface we are about to conduct a usability study, which might lead to further improvements and a qualitative comparison of *semantify.it* with other annotation tools.

1) Incoming data processing: In the future, we will put a lot of emphasis on processing of data, which is already structured, by a database or an API, but not annotated. For that we are planning on enhancing our extension system (see III-D) towards being more flexible and generic. To reach that we plan on integrating an RML processor [40] and providing templates to easily describe data sources in RML. This will improve the work flow of integrating new structured data sources a lot and help to provide more incoming data sources for *semantify.it* users.

2) Advanced validation: To improve data quality of annotations we will provide more advanced validation measures to the annotation editor and validation mechanisms to the upload and API interfaces. Based on the ideas presented in Şimşek et al. [24] we will provide a rule designer interface and a set of ready made rules to support users of *semantify.it* in generating and storing only semantically valid, high quality data.

3) Advanced analysis and reasoning: For more ambitious users of *semantify.it* we will provide an improved analytics feature. This functionality will let users see statistics about the number of annotations they made, how much classes and properties they are using and detailed statistics about how often and by whom their annotations are fetched. This will provide users with a better inside into who consumes their data and will hence lead to better annotations. And since the *semantify.it* platform stores a huge amount of public accessible, structured, high quality data we think of also setting up a overall statistics website with anonymous insides into what data is available and the performance of *semantify.it*'s user's websites. The data can maybe also be used, anonymized of course, to make predictions about certain fields in which a lot of websites use *semantify.it* for annotations. As the part of our future work, we will create a knowledge graph for tourism in Tyrol region in Austria by exporting relevant annotations from the *semantify.it* via the REST API and loading them into a triple store after preprocessing (e.g., identification, reconciliation). This will help us to apply reasoning and reveal the implicit

 ISBN: 978-1-61208-600-2

knowledge. Additionally, with the help of historical data, we will be able to apply daata analytics such as showing the price trend thoughout a year in a region.

4) On thy fly adaptability to schema.org versions: The consortium behind schema.org tries to drive development by releasing new versions of schema.org in relatively short cycles. The updates mostly feature significant changes to the core vocabulary. In version 3.1, for example, 12 classes and 10 properties for accommodation businesses were introduced (as described in [51]). To keep *semantify.it* always up do date we are going to implement an on-the-fly adaptability feature where, whenever a new schema.org version is released, the *semantify.it* editor uses the latest vocabulary from the Github repository of schema.org.

ACKNOWLEDGMENT

The authors would like to thank the Online Communications working group (OC) [52] for their active discussions and input during the OC meetings, our colleague Oleksandra Panasiuk for the creation of the domain specifications for showcase domains on *semantify.it*, our colleague Zaenal Akbar for the adaption of the Feratel wrapper, our colleague Boran Taylan Balci for the programming of the corporate blog scraper and the *semantify.it* development team (in the order of joining the team) Omar Holzknecht, Roland Gritzer, Richard Dvorsky and Dennis Sommer, for their hard work and their commitment to the mission of making the Semantic Web real and usable for everyone. A special thank you goes to all the testers and pilots who were giving *semantify.it* a chance, regardless the missing and often buggy features along the way. There is no good product without good testers!

REFERENCES

[1] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked data on the web (ldow2008)," in Proceedings of the 17th international conference on World Wide Web. ACM, 2008, pp. 1265–1266.

[2] O. Lassila and J. Hendler, "Embracing" web 3.0"," IEEE Internet Computing, vol. 11, no. 3, 2007.

[3] "The headless web," https://paul.kinlan.me/the-headless-web/, accessed: 2017-11-06.

[4] "Amazon echo," https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-Alexa-Black/dp/B00X4WHP5E, accessed: 2017-11-06.

[5] "Apple siri," https://www.apple.com/ios/siri/, accessed: 2017-11-06.

[6] "Google allo," https://allo.google.com/, accessed: 2017-11-06.

[7] "Microsoft cortana," https://www.microsoft.com/en-us/windows/cortana, accessed: 2017-11-06.

[8] R. Meusel, C. Bizer, and H. Paulheim, "A web-scale study of the adoption and evolution of the schema. org vocabulary over time," in Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. ACM, 2015, p. 15.

[9] R. V. Guha, D. Brickley, and S. Macbeth, "Schema. org: Evolution of structured data on the web," Communications of the ACM, vol. 59, no. 2, 2016, pp. 44–51.

[10] "Schema.org documentation website," http://schema.org, accessed: 2017-11-06.

[11] "Hypertext markup language (html)," https://en.wikipedia.org/wiki/HTML, accessed: 2017-11-06.

[12] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton, "Rdfa in xhtml: Syntax and processing," Recommendation, W3C, vol. 7, 2008.

[13] "Microdata at w3c," https://www.w3.org/TR/microdata/, accessed: 2017-11-06.

[14] M. Lanthaler and C. Gütl, "On using json-ld to create evolvable restful services," in Proceedings of the Third International Workshop on RESTful Design. ACM, 2012, pp. 25–32.

[15] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström, "Json-ld 1.0," W3C Recommendation, vol. 16, 2014.

[16] E. Kärle, A. Fensel, I. Toma, and D. Fensel, "Why are there more hotels in tyrol than in austria? analyzing schema. org usage in the hotel domain," in Information and Communication Technologies in Tourism 2016. Springer, 2016, pp. 99–112.

[17] "Semantify.it," https://semantify.it, accessed: 2017-11-06.

[18] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," Web Semantics: science, services and agents on the World Wide Web, vol. 4, no. 1, 2006, pp. 14–28.

[19] A. Khalili and S. Auer, "User interfaces for semantic authoring of textual content: A systematic literature review," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 22, oct 2013, pp. 1–18. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1570826813000498

[20] L. Reeve and H. Han, "Survey of semantic annotation platforms," in Proceedings of the 2005 ACM symposium on Applied computing. ACM, 2005, pp. 1634–1638.

[21] A. Gangemi, "A Comparison of Knowledge Extraction Tools for the Semantic Web." Springer, Berlin, Heidelberg, 2013, pp. 351–366. [Online]. Available: http://link.springer.com/10.1007/978-3-642-38288-8_24

[22] A. Khalili and S. Auer, "WYSIWYM Authoring of Structured Content Based on Schema.org," in Web Information Systems Engineering – WISE 2013: 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part II, X. Lin, Y. Manolopoulos, D. Srivastava, and G. Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 425–438.

[23] J. L. NavarroGalindo and J. Samos, "The FLERSA tool: adding semantics to a web content management system," International Journal of Web Information Systems, vol. 8, no. 1, mar 2012, pp. 73–126. [Online]. Available: http://www.emeraldinsight.com/doi/10.1108/17440081211222609

[24] U. Şimşek, E. Kärle, O. Holzknecht, and D. Fensel, "Domain specific semantic validation of schema.org annotations," in A.P. Ershov Informatics Conference (the PSI Conference Series, 11th edition). Springer (To appear), 2017. [Online]. Available: http://arxiv.org/abs/1706.06384

[25] "W3c web annotation data model," https://www.w3.org/TR/annotation-model/, accessed: 2017-11-06.

[26] "W3c web annotation vocabulary," https://www.w3.org/TR/annotation-vocab/, accessed: 2017-11-06.

[27] "Google introduction to structured data," https://developers.google.com/search/docs/guides/intro-structured-data#markup-formats-and-placement, accessed: 2017-11-06.

[28] W. W. W. Consortium et al., "Json-ld 1.0: a json-based serialization for linked data," 2014.

[29] "Wordpress," https://wordpress.com/, accessed: 2017-11-06.

[30] "Typo3," https://typo3.org/, accessed: 2017-11-06.

[31] "Joomla," https://www.joomla.org/, accessed: 2017-11-06.

[32] "Drupal," https://www.drupal.org/, accessed: 2017-11-06.

[33] "Semantify.it bitbucket repository," https://bitbucket.org/semantifyit, accessed: 2017-11-06.

[34] "Semantify.it github repository," https://github.com/semantifyit, accessed: 2017-11-06.

[35] Z. Akbar, E. Kärle, O. Panasiuk, U. Şimşek, I. Toma, and D. Fensel, "Complete semantics to empower touristic service providers," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, 2017, pp. 353–370.

[36] P. Buxmann, T. Hess, and S. Lehmann, "Software as a service," Wirtschaftsinformatik, vol. 50, no. 6, 2008, pp. 500–503.

[37] "Git," https://git-scm.com/, accessed: 2017-11-06.

[38] "Material design," https://material.io/, accessed: 2017-11-06.

[39] https://w3techs.com/technologies/history_overview/content_management, accessed: 2017-11-06.

 ISBN: 978-1-61208-600-2

[40] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle, "Rml: A generic language for integrated rdf mappings of heterogeneous data." in LDOW, 2014.

[41] "Uuid wikipedia," https://en.wikipedia.org/wiki/Universally_unique_identifier, accessed: 2017-11-06.

[42] "Tvb mayrhofen," http://mayrhofen.at, accessed: 2017-11-06.

[43] "Tvb seefeld," http://seefeld.com, accessed: 2017-11-06.

[44] "Tvb fügen," http://best-of-zillertal.at, accessed: 2017-11-06.

[45] "Tirolwerbung.at," http://tirolwerbung.at, accessed: 2017-11-06.

[46] "Semantify.it wordpress plugin," https://wordpress.org/plugins/semantify-it/, accessed: 2017-11-06.

[47] "Semantify.it typo3 plugin," https://typo3.org/extensions/repository/view/semantify_it, accessed: 2017-11-06.

[48] "Cloudflare," https://www.cloudflare.com/, accessed: 2017-11-06.

[49] "Pingdom website speed test," https://tools.pingdom.com, accessed: 2017-11-06.

[50] "Feratel web client," http://www.feratel.at/loesungen/feratel-destination/datenmanagement/web-client/, accessed: 2017-11-06.

[51] E. Kärle, U. Simsek, Z. Akbar, M. Hepp, and D. Fensel, "Extending the schema. org vocabulary for more expressive accommodation annotations," in Information and Communication Technologies in Tourism 2017. Springer, 2017, pp. 31–41.

[52] "Sti innsbruck," http://oc.sti2.at/, accessed: 2017-11-06.

 ISBN: 978-1-61208-600-2

Hunting for Direct Translations across Wikipedia Articles

Duc Manh Hoang and Marco Ronchetti
Department of Information Engineering and Computer Science
Università di Trento
Povo di Trento, Italy
email:ducmanh.hoang@studenti.unitn.it, marco.ronchetti@unitn.it

***Abstract*—This paper deals with a problem in the area of Cross-Language Plagiarism Detection. In particular, it presents a system, able to detect portions of a Wikipedia page, which have been obtained by translating a Wikipedia page on the same semantic content but written in a different language. The problem is relevant in the context of Wikipedia pages maintenance, and could be of interest in other areas such as news comparison in different languages. We discuss the problem, the system and its implementation and briefly present its evaluation.**

Keywords-Machine translation; Wikipedia; Plagiarism.

I. INTRODUCTION

Although the term "Encyclopedia" was first used in the XVI-th century, the most important attempt to code all the mankind knowledge under such name happened in the mid of the XVIII century. It was coordinated by Denis Diderot and had a very important influence on the development of the Age of Enlightenment, which shaped the western culture for the years to come [1]. In our lives, we witnessed another cultural revolution connected with the Encyclopedia concept: the birth and growth of Wikipedia, the largest cultural collaborative writing effort in mankind history. The importance of Wikipedia cannot be overstated. According to Alexa [2], it ranks 5th among the most visited Web sites, and is the first non-commercial one, being surpassed only by Google (www.google.com), YouTube (www.google.com), Facebook (www.facebook.com), and Baidu (www.baidu.com).

Semantic information can be extracted from Wikipedia: the DBPedia initiative pioneered such effort [3], allowing to build semantic applications on top of it (see, e.g., [4-6]).

But Wikipedia is not just "one" collaborative Encyclopedia. It is rather a collection of many versions in different languages: presently 295 (but 10 do not reach 100 pages). The English Wikipedia contains more than 5 million articles, while versions in 12 other languages exceed 1 million articles, and 124 more languages contain at least 10.000 entries [7]. Recognizing the importance of the multilingualism, Wikipedia offers special links among pages dealing with the same topic, but written in different languages: the so-called "interlinks". Interlinks allow users to easily browse the corresponding pages in other languages, and hence to compare and integrate the knowledge contained in a page with the one of the other Wikipedia versions (provided the different language is not a hurdle for the curious reader). In fact, it is quite natural that some entries are richer in a language that in another, as this reflects a "national interest". For instance, a type of German locomotive not having a special historical value is (probably well) documented in the German Wikipedia, but hardly mentioned in other languages or, when mentioned, the article in languages other than German will probably be sketchy and contain only some of the most important details. In spite of this, non-German railway historians will nonetheless be interested in finding out more.

A comparison of pages in different languages is also useful for editors, who wish to integrate a page in a language, gathering knowledge via the interlink.

In an attempt to increase the number of pages (especially for languages with a limited coverage), Wikipedia has been recently promoting the translation of pages, which exist in the "main editions" and are absent in other languages. Since there exists a procedure for acknowledging that a page has been translated [8], such form of "plagiarism" has no negative connotation.

It is interesting for various reasons to find out if a Wikipedia page in a given language has been (partially or totally) translated in other languages. We therefore asked ourselves, if there is a way to automatically detect such translations. For instance, it could help identifying semantic difference between papers in different languages (e.g, missing parts) and could be used to automatically signal the necessity or opportunity to improve a page in a given language.

We developed a software tool to deal with such problem. In the present paper, we describe its architecture and working mechanism, and present a sample of the results obtainable with it. Section II presents the relation of our work with the area of Cross-Lingual Plagiarism Detection; Section III discusses how we decided to attack the problem of comparing Wikipedia pages on a given topic, written in different languages; Section IV describes the process of comparing the pages; Section V presents the overall software architecture. In Section VI we briefly presents the evaluation, and finally in Section VII we draw our conclusions.

II. RELATION WITH CROSS-LINGUAL PLAGIARISM DETECTION

Our problem has some common traits with Cross-Lingual Plagiarism Detection (CLPD), which has been studied by several authors (see, e.g., [9-12]). There are, however,

 ISBN: 978-1-61208-600-2

differences. Plagiarism Detection aims at finding whether a "suspect" document contains parts of text taken by any document on the Web. In the multilingual case, the problem is exacerbated by the difficulty of finding a set of candidate sources written in other languages, so that the simple strategy of using traditional search engines is not effective. Then, also the comparison of suspect and potential source is more difficult since it has to be performed across languages.

A typical architecture for CLPD [10] comprises heuristic retrieval (i.e., the gathering of possible sources), detailed analysis (to compare the suspect with every document collected by the retrieval) and heuristic post-processing (for merging or discarding possible sources).

Our case is simpler, since our set is predefined by semantics, and it is the set of documents related by interlinks. We can, hence, focus on the second part of the problem, avoiding heuristic retrieval, and have fewer difficulties in dealing with it.

Also, there is another important difference: plagiarism is usually considered as an unacceptable practice. Plagiarists hence often try to disguise the copied parts, e.g., by paraphrasing portions of the text, so as not to be detected by search engines. Instead, in the case we are interested in, copying is a socially accepted and even encouraged practice, which helps spreading the knowledge to other communities, and therefore authors do not need to try to hide it.

Typical strategies for Cross Language analysis include lexicon-based systems, thesaurus-based systems, comparable corpus-based systems, parallel corpus-based systems and machine translation-based systems. We cannot discuss all of them here, as the area is wide, and refer the reader to [13]. The approach we chose, which is machine translation-based, is described in the following sections.

III. CONSIDERING WIKIPEDIA PAGES WRITTEN IN DIFFERENT LANGUAGES: THE NORMALIZATION PROCESS

The problem to solve is to be able to compare a pair of corresponding pages in different versions (i.e., languages) of Wikipedia: say P^x_Y and P^x_z (Page X in Language Y and Page X in Language Z). The way to compare a pair of pages could be to try to extract the contained semantic information, mapping it to an ontology and comparing it. We think that such an approach is bound to fail. In fact, since both P^x_Y and P^x_z are about x, the semantic meaning obviously matches. The richness of the semantics could be different (as one could contain more details than the other), but even if the richness is the same, this does not imply that a page is the translation of the other. More information about the structure and actual content of the page has to be taken into account.

Such information must come from the texts we want to compare, but they are in different languages. To make them comparable, we decided to translate them. In order to make our approach scalable, we opted to use automatic translation. We were well aware of the limits that today's machine translation (MT) has, but decided anyway to give it a try to verify if, in spite of them, the approach could work.

Having to compare a German and a French page on the same topic (P^x_G and P^x_F) we could decide to translate one of them in the other language, and then compare say $P^x_{G->F}$ and P^x_F, where the suffix X->Y means page "written in language X and translated into language Y". This introduces an asymmetry, so we could also compare P^x_G and $P^x_{F->G}$ and then match the two results.

However, we thought that such an approach would have presented some problems. First, we were interested in checking not only two languages, but a set of the largest Wikipedia versions (namely, we chose English, French German and Italian). This would have implied multiple translations. Second, the quality of publicly and freely available MT engines seems far from being uniform when translating between languages. Since the technology used by engines, such as Google Translate is considered a trade secret, it is difficult to find evidence in academic papers on what is going on behind the scenes. There are of course reviews of MT techniques (such as, e.g., [14]), and indications that Google uses "mostly" statistical methods [15], which make unnecessary to "bridge" though an intermediate language or model. In any case, the quality of translation into English seems to be better than the one into other target languages, maybe also because its grammar is far simpler that the one of many other languages, including the ones we have chosen for our exercise, or because of a larger base, since English is today's *lingua franca*. We cannot prove this assumption, as we did not find scientific evidence for this fact. However, combining the combinatorial problem with the guess that translation into English is at least not worse than translations into other languages, we decided to "normalize" all the texts (written in other than English languages) by translating them into English. Hence, for every topic X we are interested in, we consider the set $\{P^x_E, P^x_{F->E}, P^x_{G->E}, P^x_{I->E}\}$.

We therefore wrote a software component which, given a Wikipedia page in one of the four languages, checks if the interlinks into the other three languages are present, and once they are found it performs the needed translation. We could use several MT engines (Bing, Google, SDL, Yandex). According to evaluations available on the Web, they seem to provide similar performances. Once again, we were facing the impossibility to base our work on scientifically sound grounds, but had to trust information which, in spite of being rather coherent, does not offer scientific rigor. In the end we decided to use the Yandex API [16] to perform the translation, since they were the most inexpensive available option (with up to 2 million characters/month free, and the cheapest option above that threshold).

IV. COMPARING THE PAGES

For a given topic X, we now have four documents: P^x_E, $P^x_{F->E}$, $P^x_{G->E}$, $P^x_{I->E}$. To compare the pages, we first segment the text by breaking their content into sentences. We use a list of abbreviation to avoid getting confused by the punctuation used for abbreviations rather than for ending sentences.

The next step is to apply to each sentence N-Gram segmentation, a technique for breaking a stream of text into units of N ordered adjacent words [17]. Part–of-speech (P.O.S.) tagging is then applied to identify the role of each word in the sentence (e.g., noun, verb, adjective etc.). P.O.S.

 ISBN: 978-1-61208-600-2

tagging is needed to perform the next operation, which is lemmatization: a technique similar to stemming but aware of the context in which a word is situated. This allows replacing, e.g., "better" with "good", verbs (such as I "am") into their infinitive form ("be"), etc. Stop words (such as articles, but also any very frequent word) are then removed: since they are very common, their presence in unrelated sentences would generate noise in terms of false positives when comparing their content, so it is better not to have them in the text (even if by doing so some relevant part of "meaning" gets omitted).

At this point, each of the four normalized documents have been exploded in a set of cleaned-up sets of words $\{S^x_{Li}\}$, where L stands for the original language (although all documents now contain only English words) and i is the index of the phrase in the document. The documents P^x_L, which are at the origin of our sets, generally have different number of sentences, which we will call N^x_L, so for each S^x_{Li} the index i runs from 1 to N^x_L.

Let us now try to ascertain that a portion of document P^x_A has been copy-translated into P^x_B, or vice versa. We can examine pair or sentences, but we cannot make assumptions on where they are: a portion from the beginning of a document could have been copied onto the central part of the other, so we need to compare each sentence in P^x_A with every other sentence in P^x_B. This will generate a matrix of dimension N^x_A x N^x_B, in which the cell (i,j) contains a number representing a measure of similarity between the sentences S^x_{Ai} and S^x_{Bj}.

We now need to know how such measure is computed, and what can we do with the matrix.

To evaluate sentence similarity, we tested two different approaches: we used both Cosine similarity [18] and Jaccard similarity [19]. For each pair of sentences $\{S^x_{Ai}, S^x_{Bj}\}$ we build a bag of words, containing all the words which appear in at least one of the two sentences (but each word is present only once in the bag, regardless of the actual number of occurrences in the sentences). The words are ordered (in an arbitrary way), defining in this way an M-dimensional space, where M is the cardinality of the bag of words. For each sentence, we can then compute its position in such vector space: the number of occurrences of the z-th word in it gives the value of the z coordinate. Having the coordinates of the two sentences, their Cosine similarity is evaluated as the scalar product between the vectors, which represent them (such value is in the interval [0,1], since only the positive subspace is considered, as the number of occurrences which determine the coordinates cannot be negative).

The Jaccard Similarity is instead computed as the ratio between the cardinality of two sets: $|S^x_{Ai} \cap S^x_{Bj}| / |S^x_{Ai} \cup S^x_{Bj}|$. This value is also in the interval [0,1].

At this point we forked our project, using these two different measures of similarity (Cosine and Jaccard) and then proceeding in the same way. In both cases, we end up with a score matrix for topic X and the pair of languages {A,B}, and in both cases the values of the cells in the matrix are numbers between 0 and 1.

The closest a cell is to one, the highest the similarity between the two corresponding phrases. However, in the Wikipedia page generation case, a "copy-translate" is not just related to one single sentence, but rather to a section of the paper, which consists of multiple adjacent sentences, each with a high similarity value. Hence we are interested in detecting diagonal subsets with high similarity values in the score matrix. For instance, we are interested in finding situations where not only S^x_{Ai} and S^x_{Bj} are similar, but also the pairs $\{S^x_{Ai+1}, S^x_{Bj+1}\}$, $\{S^x_{Ai+2}, S^x_{Bj+2}\}$, …, $\{S^x_{Ai+n}, S^x_{Bj+n}\}$.

To facilitate the identification of such sequences, we canceled the noise, by putting to 0 all the cells, which have a value less than a given threshold. To define the threshold level, we considered how the data are distributed in the matrix, and assumed a Gaussian distribution for the noise. We kept only the tail of the high values. We then looked for diagonals sequences: these reveal portions of the text, which are very similar and hence are likely to be copy-translated.

V. OVERALL ARCHITECTURE

We summarize the architecture of our system, which is outlined in Figure 1. A harvester (Document Reader) gets the documents and generates a set composed by a given Wikipedia page in one of the four languages and the interlinked pages in the other three languages.

It then translates all the non-English pages into English, obtaining a set of quadruples $\{P^x_E, P^x_{F->E}, P^x_{G->E}, P^x_{I->E}\}$. More details about the Document Reader are given later.

Given the quadruple, each of its documents is passed to the Preprocessing Unit, which performs text segmentation (into phrases), P.O.S. tagging, lemmatization and stop words removal.

The result is given to the Text-Similarity Unit, which evaluates Cosine similarity and Jaccard similarity between each pair of sentences contained in each pair of elements of the quadruple.

The output of the Text-Similarity Unit is passed to the Post Processing Module. Here, for each pair (P^x_A, P^x_B) where A and B are two different elements of the set {E, F->E, G->E, I->E), the values computed by the text-similarity unit compose the score matrix, which is cleaned discarding the low values, and for which non-null diagonal sequences are searched. If some diagonals are found, we have a candidate "copied" section of the articles. Of course, similarity is symmetric, so we still need to know which is the original, and which the copy. This can be easily understood by checking the version in Wikipedia history. The task is hence accomplished, and we can pass to the evaluation phase.

 ISBN: 978-1-61208-600-2

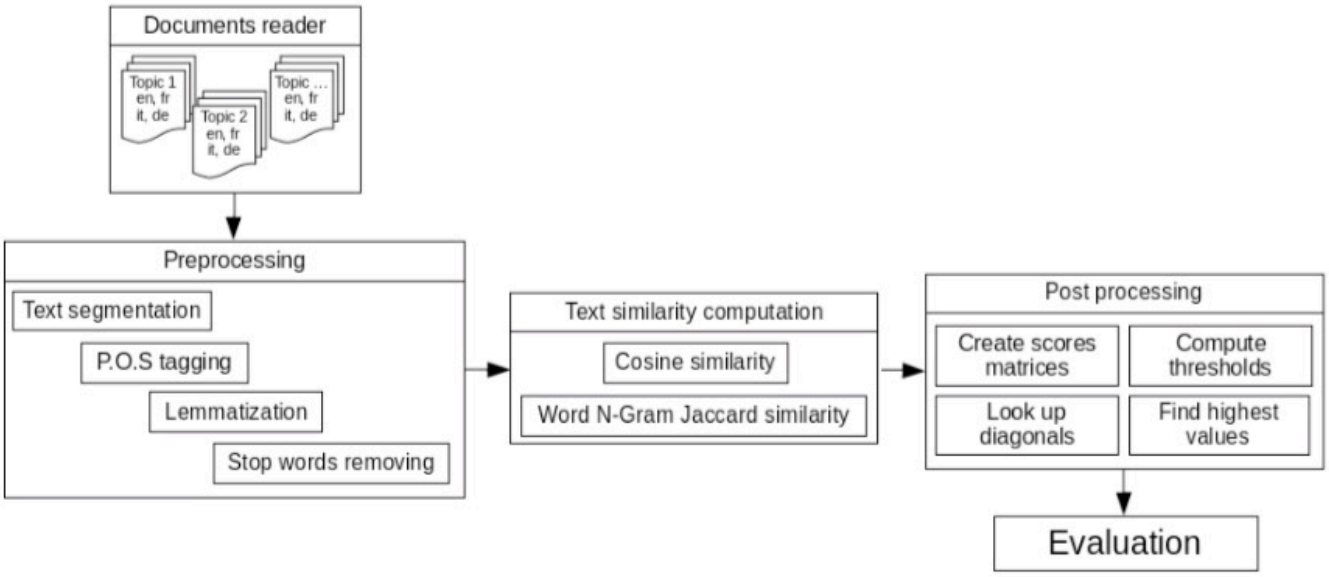

Figure 1. Overall logical architecture of the system (see text for a description).

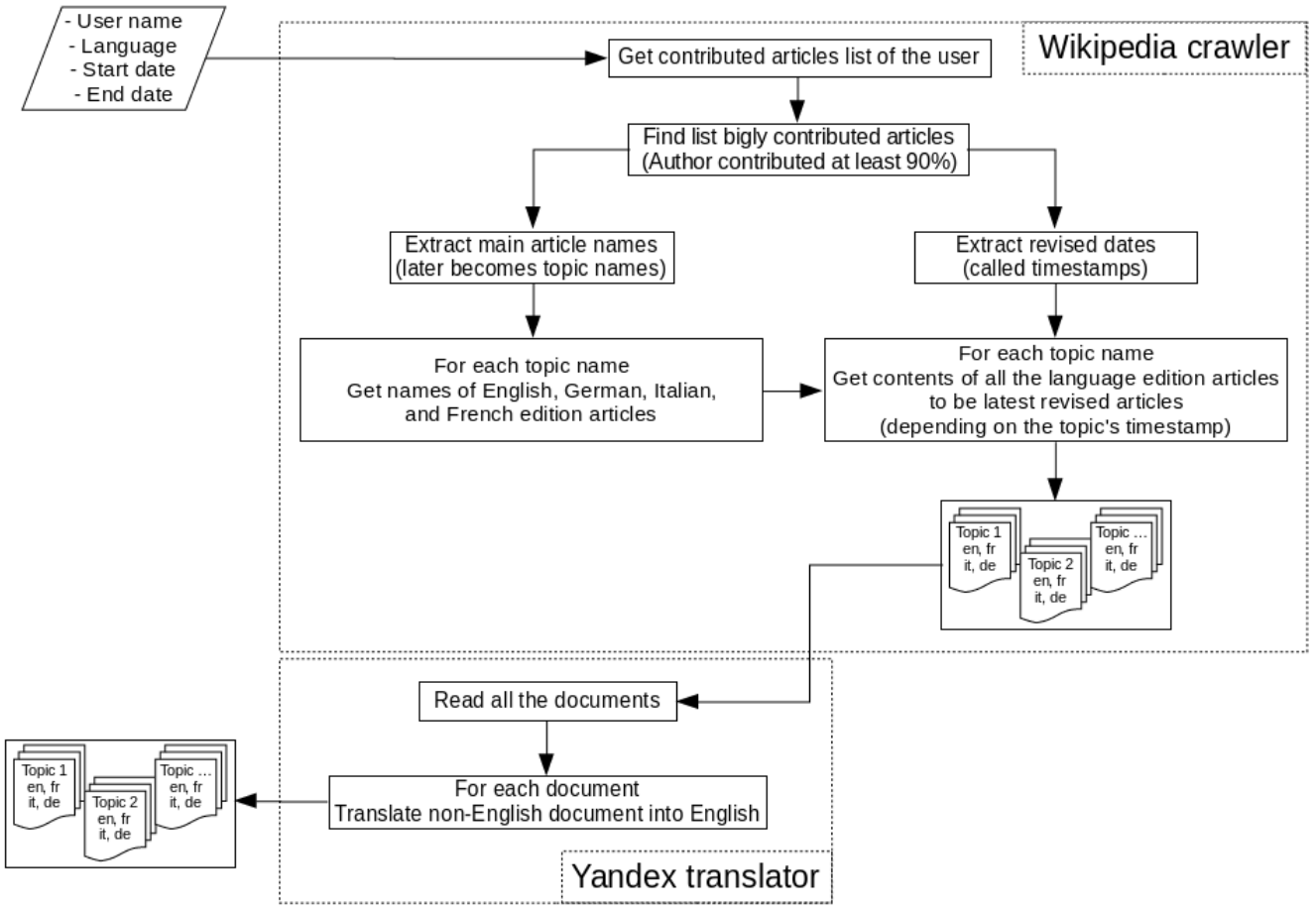

Figure 2. Explosion of the Documents Reader component (see text for a description).

Let us now come to a more detailed description of the harvester (Document Reader), which is exploded in Figure 2. It is composed by the Wikipedia Crawler and a Translation Unit. The Wikipedia Crawler is fed with the specification of the user name of a Wikipedia author, a language and a time span running from a start date to an end date. The Crawler extracts the list of all the Wikipedia articles (in the given language) that the user contributed to in the selected time span, and selects those for which the author was the major contributor (according to a customizable percentage parameter). For these, the interlinked articles are retrieved. Article revisions are considered to make sure that the compared versions refer to the same time. This is very important, since an article, which was used as a source for a copy-translation, could have been modified after the translation was performed.

Once the set of four documents has been generated, it is necessary to translate the Italian, French and German ones into English. In our implementation this is done with the Yandex translator, but the Translation Unit could use any other translator.

The main software tools we use are the Wikipedia API [20], DKPro Core Library [21], the UIMA-Unstructured Information Management Architecture [22] and the already mentioned Yandex API.

VI. EVALUATION

Evaluating the results has been a costly operation, since we need an "oracle" able to give us a human evaluation of whether a portion of a page has been copy-translated. The time needed to perform such an operation is non-negligible, and the results are not always clear-cut. Sometimes a portion of the document is not simply translated, but reworked and paraphrased. In other cases, semantic identity of the content pushes the authors to write very similar sentences, even when being unaware of each-other's work: one has to remember that the topic of the considered pages in a given set is the same, and hence what is classified ad "paraphrase" could well simply be due to "semantic similarity".

We used 56 topics containing 148 pages generated by 4 authors. Author 1 translated some pages from Italian Wikipedia to the French one. Author 2 translated from German Wikipedia to the Italian one. Author 3 translated from English to French. Author 4 did not rely his/her contributions on copy-translation.

Our (human) judgment of these authors is reported in Table 1.

TABLE I. EVALUATION SET

Author	Total pages	Partially copied pages	Paraphrased pages
Author 1	33	7	10
Author 2	30	10	4
Author 3	59	21	5
Author 4	26	0	5

For each pair P^x_A, P^x_B, we manually evaluated whether the version of the pages in other languages had significant, similar sections (we will call this "oracle evaluation"). We categorized the pairs of pages (a topic in two languages) by using three descriptors: copied, paraphrased, not copied (where a page is considered to be "copied" if a significant section of it (at least 4 or 5 sentences) is similar to a page written in a different language.

We then compared the human annotated results with the predictions of our system, and checked if there was a full agreement (both systems stating the same thing), partial agreement (the machine declaring that there was a copy, and the human describing the mapping as a paraphrase) or no agreement (oracle and machine producing opposite statements). The possible cases and the corresponding results are reported in Table 2.

TABLE II. EVALUATION RESULTS

Oracle	Prediction	Evaluation	Numerosity
YES	YES	True positive	31
YES	NO	False negative	10
NO	YES	False positive	1
NO	NO	True negative	82
Paraphrase	YES	Uncertain	7
Paraphrase	NO	Uncertain	17

In 16% of the cases we examined, the oracle was uncertain whether there had been a copy-translation between the considered pair of documents.

Out of the cases where the oracle decided with certainty for the NO, the system prediction was right 99% of times .
Out of those where the oracle decide with certainty for the YES, the system prediction was right 75% of times.

Seen from a different perspective and taking into account also the cases when the oracle was uncertain, whenever the system predicted the presence of copy-translation, it was right 79% of times. When it predicted its absence, it was right 75% of times.

We find no difference in using the matrices obtained using Cosine similarity and Jaccard similarity: both measures yield results of the same quality.

VII. DISCUSSION AND CONCLUSIONS

We presented our work on finding whether a Wikipedia page originated from another one, written on the same topic but in a different language, by translating a portion of the page. The work is somehow close to the domain of Cross-Language Plagiarism Detection, but presents some peculiarity, which distinguishes it from the mainstream in that area.

The work can be the basis for tools, which could be useful for Wikipedia maintainers, and could be used for statistical analysis of the Wikipedia body of knowledge. For instance this work, given a Wikipedia author, could help classifying her/his type of contributions.

The evaluation of the system we developed shows a very good reliability in a domain, where even humans have difficulty to establish with certainty the truth.

 ISBN: 978-1-61208-600-2

In future, it would be interesting to examine if our approach also works with other languages, such as the Asian ones.

The developed software has been released in public domain and is publicly available at [23]. Some more detailed explanations are available there in the readme file, which also reports a sample of the experiment. For any additional clarification, interested people are invited to contact the authors.

REFERENCES

[1] P. Blom, "Enlightening the world: Encyclopédie, the book that changed the course of history", New York: Palgrave Macmillan (2005)

[2] https://www.alexa.com/topsites, last visited Sept. 6, 2017

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a Web of open data." The semantic Web (2007): pp. 722-735.

[4] A. Prato, and M. Ronchetti, "Using Wikipedia as a reference for extracting semantic information from a text." Advances in Semantic Processing, 2009. SEMAPRO'09. Third International Conference on. IEEE, 2009. pp. 56-61

[5] F. Valsecchi and M. Ronchetti, "Spacetime: a two dimensions search and visualisation engine based on linked data." Conference on Advances in Semantic Processing (SEMAPRO). 2014.

[6] M. Battan and M. Ronchetti, "QwwwQ: Querying Wikipedia Without Writing Queries." International Conference on Web Engineering. Springer, Cham, 2016. pp. 389-396

[7] https://www.wikipedia.org/ last visited Sept. 6, 2017

[8] https://en.wikipedia.org/wiki/Wikipedia:Translation, last visited Sept. 6, 2017

[9] C. K. Kent and N. Salim, "Web based cross language plagiarism detection." Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on. IEEE, 2010. pp. 199-204

[10] A. Barrón-Cedeño, P. Gupta, and P. Rosso, "Methods for cross-language plagiarism detection." Knowledge-Based Systems 50 pp. 211-217, 2013.

[11] M. Franco-Salvador, P. Gupta, and P. Rosso, "Cross-language plagiarism detection using a multilingual semantic network." European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2013. pp. 710-713

[12] E. Nava, F. Wm. Tompa, and A. Shakery, "Using a dictionary and n-gram alignment to improve fine-grained cross-language plagiarism detection." Proceedings of the 2016 ACM Symposium on Document Engineering. ACM, 2016. pp. 59-68

[13] M. Potthast, A., Barrón Cedeño, B., Stein, and P. Rosso, "Cross-language plagiarism detection. Language Resources and Evaluation (LRE)", Special Issue on Plagiarism and Authorship Analysis 45 (1), pp. 1–18. 2011.

[14] S. Tripathi and J. K. Sarkhel, "Approaches to Machine Translation", Annals of Library and Information Studies, Vol 57, pp. 388-393, 2010.

[15] T. Mikolov, V. Le Quoc, and I. Sutskever, "Exploiting similarities among languages for machine translation." arXiv preprint arXiv:1309.4168 (2013).

[16] https://tech.yandex.com/translate/, last visited Sept. 6, 2017

[17] P. Norvig, "Natural language corpus data". In Beautiful Data, edited by T.Segaranand and J.Hammerbacher, pp. 219–242. Sebastopol, Calif.: O'Reilly (2009)

[18] A. Singhal, "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43 (2001)

[19] P. Jaccard, "The distribution of the flora in the alpine zone", New Phytologist, 11: 37–50 (1912)

[20] https://www.mediawiki.org/wiki/API:Main_page, last visited Sept. 6, 2017

[21] https://dkpro.github.io/dkpro-core/, last visited Sept. 6, 2017

[22] https://uima.apache.org/, last visited Sept. 6, 2017

[23] https://github.com/ducmanhhoang/Wikipedia-Matching, last visited Sept. 6, 2017

 ISBN: 978-1-61208-600-2

Use of Negation Markers in German Customer Reviews

Amelie I. Metzmacher[a], Verena Heinrichs[a], Björn Falk[a], Robert H. Schmitt[a]
[a] Laboratory for Machine Tools and Production Engineering (WZL), RWTH Aachen University, Aachen, Germany
E-mail: A.Metzmacher@wzl.rwth-aachen.de, V.Heinrichs@wzl.rwth-aachen.de, B.Falk@wzl.rwth-aachen.de, R.Schmitt@wzl.rwth-aachen.de

***Abstract*—The research presented examines the use of negation markers in German customer reviews. The objective is to identify differences, as well as similarities in the use of language for reviews rating material products and services. Therefore, in an annotation study, negation markers and sentiment values of customer reviews rating these product categories had to be assigned. The results obtained confirm the hypothesis that customer reviews relating to services contain more negation markers than customer reviews rating material products. However, there exists no significant difference in token distances between the negation marker and the sentiment decisive part of speech (POS). Finally, the findings should be applied in a machine learning algorithm for extracting relevant information from German customer reviews.**

Keywords-Social Media; Customer Reviews; NLP; Sentiment Analysis.

I. INTRODUCTION

More than 80% of all customers express their experiences with products and services in social media [1]. Being publicly available, information provided by customer reviews in social media is important for both potential customers and companies. 80% of all potential customers base their purchase decision on the experiences of other customers trusting the judgment of strangers more than the recommendations of family and friends [1]. For companies, customer reviews provide insights with respect to not only the product, its functionalities or the service experience but also regarding the person who is the customer, his needs, wishes and persona. In analyzing customer feedback, companies are able to gain up-to-date and authentic knowledge about the product, the service and the customer.

The steadily growing number of customer reviews available in social media requires text mining and machine learning techniques such as sentiment analysis for a detailed understanding of the information provided by customer reviews. A major issue in sentiment analysis is the identification and handling of negation markers [2] - [6]. Negation markers cause valence shifts in customer reviews, e.g., shifting a positive statement into a negative statement and vice versa [7]. Since the use of negation markers is often very language-specific, language dependent approaches and algorithms are needed to analyze the sentiment of customer reviews correctly.

The German language shows a strong tendency for conventional indirectness by using syntactic downgraders. Syntactic downgraders modify the intended illocutionary act, i.e., the meaning conveyed, of the speaker towards the audience [8]. Examples of syntactic downgraders are modal verbs, tense or negation markers [9].

There exist different types of negation markers relating to different language levels. Examples given are firstly listed in the English translation and secondly in the German original word. On the morphological level, negation is expressed using distinct prefixes (e.g., “unhappy”, “unglücklich”) or suffixes (e.g., “senseless”, “sinnlos”). Discrete meaning bearing units in place of morphemes (e.g., “not”, “nicht”) represent the class of syntactic negations. In addition, the use of diminishers (e.g., “hardly”, “kaum”) negates the inherent polarity of an expression [2] [3] [9]. Negation markers either apply to words directly carrying a certain sentiment (local negation) or relate to words that do not carry a sentiment (long distance negation; indirect) [4].

Following Giora et al. [10], negation markers function as an instruction from a speaker to an audience to suppress the negated information. By using negation markers to suppress the negated information for the hearer, the speaker saves his face and remains polite when communicating a negative statement [11]. Politeness is a distinct feature of civilized societies and thus seen as an important social value guiding social interactions [12].

As opposed to evaluating a material product during application, the evaluation of a service includes the evaluation of the person of the service provider during service provision. The customer’s evaluation then applies directly to the professional and personal behavior of the service provider throughout social interaction.

Considering the aspect of politeness and keeping one`s face in social interaction, we hypothesize that customer reviews relating to services contain more negation markers than customer reviews rating material products. Finding significant differences might ameliorate algorithms for extracting relevant information out of social media content. In addition, it could help to identify automatically if a customer judges a product or a service. In particular, this would be of interest for companies providing both services and products.

 ISBN: 978-1-61208-600-2

To examine the above stated hypothesis, an annotation study was conducted. The database consisted of 3,767 German customer reviews in total extracted from different social media platforms. The customer reviews related to material products, as well as services. Three subjects were asked to annotate randomly chosen sentences of the reviews. The subjects' task was to annotate the sentiment of each given sentence and explicitly reference whether there is a negation marker present or not.

In the following, section 2 deals with related literature covering similar topics. In section 3, the annotation study is described, whereas in section 4 the results are analyzed and discussed. Finally, section 5 finishes this contribution with a brief conclusion and information on future working steps.

II. Related Work

Related work for the presented approach can be divided into at least two research areas: sentiment analysis as part of natural language processing and linguistic research analyzing customer reviews in social media with respect to material products and services. However, there are often overlaps, particularly if the objective has a clear focus on negations. Moreover, there are only very few contributions dealing with German customer reviews in general.

Wiegand et al. [13] give an overview about the role of negations, as well as about different approaches to include negations into sentiment analysis. They state that although integrating negations is very difficult, contributions dealing with this topic generally agreed on its high relevance for sentiment analysis. For instance, Kennedy and Inkpen [7] point out that considering the effects of valence shifters has a generally positive effect on all classification methods for reviews.

Furthermore, Asmi & Ishaya [5] integrate negation calculation rules into the general framework of a rule-based polarity classification. Therein, they defined these rules based on part of speech (POS). One main finding indicates that most negation words are classified as adverbs, suffixes, prefixes or verbs. Using this information, a dependency tree is developed. Its output is the scope of negation, which indicates how negation is interacting with other words in the sentence. Although, their approach already improved the polarity classification as there was a strong correlation between the classification results of the algorithm and those of humans, the authors point out the importance of additionally implementing prepositional negations.

In a rather current contribution, Diamantini et al. [3] apply a dependency-based parse tree to investigate the scope of negation. Implementing the negation handling component just before the sentiment calculation, meaning after all other pre-processing steps have been conducted, increases the accuracy from 64.4% to 67%. In their approach, they also regard a three-class-problem as the sentiment is distinguished between negative, positive and neutral. The authors hypothesize that those samples calculated wrong imply irony. Thus, as a necessity for future work, they suggest to extend their system to consider effects of irony.

In summary, independently from the chosen method, in most cases integrating a negation model should ameliorate the accuracy of the sentiment analysis. Concerning the use of negations in German, Wiegand et al. [13] point out the requirement for more complex processing as the negated expression either precedes or follows the actual statement. Therefore, not all findings discovered e.g., in English texts, can be applied in German texts, at least not without adjustments.

As Wiegand et al. [13] point out, not all negations indicate a negative sentiment. Thus, it is important to use syntactic knowledge and regard the context. Most approaches dealing with sentiment analysis, use reviews, which are not domain-independent, e.g., a collection of reviews of several products found on google.com or movie reviews [14]. Within the group working with corpora built from movie reviews, the classification usually follows the star ratings of the authors of these reviews [15].

Although, there are some contributions, which conducted annotation studies to produce corpora from German customer reviews, they usually aim at sentiment analysis in general and do not consider the usage of negation markers. Moreover, many do not address different domains, and in particular services, as well as material goods. For instance, Boland et al. [16] conducted a study, which focuses on different domains, but does not address the use of negations.

In summary, no study has yet been conducted in which a text corpus of domain-independent customer reviews in German is annotated with regard to sentiment and particularly negations.

A prior study by the authors indicates an influence of personal commitment on customers' writing styles while formulating a product review. This is particularly the case while rating services. On the one hand, it seems that the writing becomes more precise [17]. On the other, one might argue that the human interaction required in services leads to more polite formulations.

Thus, based on the literature review and this prior study, an investigation of the amount of negations in services compared to material products is intended. Thereby, we look for hints for the application of a more polite form of criticism within German service reviews.

III. Annotation Study

The objective of the study was to examine whether customer reviews relating to services contain more negation markers than customer reviews rating material products. The products selected are accessible to the German end-consumer. The two classes contain three product types each. For products a shoe, a hazelnut spread, and a smartphone were chosen, whereas the services contained a hotel, a financial service for online businesses, and a car service station with several stations across Germany.

Altogether 38 different social media platforms, including German discussion forums and shopping sites with user comments, were chosen as data sources. To this end, 3,767 German customer reviews relating to both, material products and services, have been extracted with the open-source Java library jsoup [18]. Prior to annotation, the reviews were parsed into single sentences using the Stanford Parser [19]. The annotation was carried out on sentence level. 1,200

 ISBN: 978-1-61208-600-2

sentences, 600 for products and services each, were randomly chosen for annotation and annotated by three subjects. The subjects were German native-speakers and familiar with the process of annotating. Each subject had to annotate 200 sentences, whereby each sentence was annotated by three subjects.

Subjects were asked to identify the sentiment of each sentence while assigning the POS, which induces the negativity, positivity or neutrality of the given statements, i.e., the level of sentiment. Herein, the opinionated words are called attributes. Moreover, the subjects were asked to determine negation markers if present in the sentence, e.g., mark the indefinite pronoun "no" ("kein") or the particle "not" ("nicht"). In addition, the negation markers were assigned to the attribute the negation is associated with. For instance, the subjects had to indicate that the negation "not" ("nicht") is associated with the attribute "good" ("gut"). Thereby, it was possible to mark more than one attribute, aspect and/or negation marker per sentence, e.g., if a conjunction was present.

The annotation process was explained to the subjects with three exemplary sentences. The sentences were chosen to show different characteristics, which influence the grade of simplicity or complexity of identifying the sentiment of a sentence and its possible negation marker. For instance, one sample contained two attributes in one sentences ("The shoe looks nice, but is too heavy") or another one included only an implicit product review, meaning an attribute without an aspect ("Too heavy."). The examples ensure that the annotation process was carried out consistently.

IV. Results And Discussion

First, the interrater reliability was investigated. Fleiss' Kappa values for assigning aspects, attributes, sentiment, and negation markers between 0.5 and 0.8 are located within a moderate level of agreement [20].

The observed frequencies of the labelled negation markers were displayed and analyzed for material products and services. The frequencies of negation markers were computed based on statements within a sentence, i.e., based on attributes within a sentence.

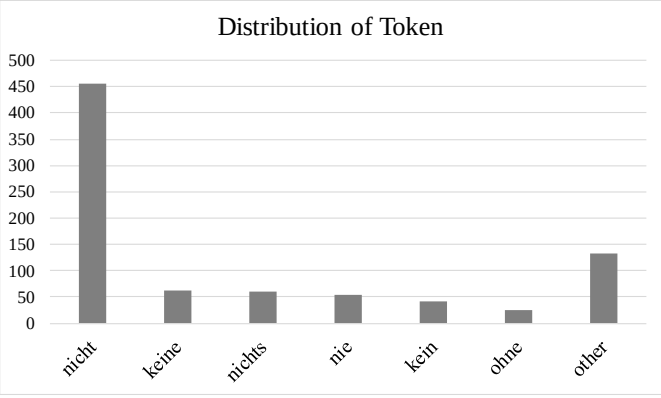

Figure 1: Distribution of token used as negation markers on sentence level

Figure 1 depicts the most used negation markers for both categories. The most frequent negation marker is the participle "not" ("nicht"). In comparison with the other frequently appearing negation markers, the participle "not" ("nicht") is not as harsh as the negation marker "any" ("keine"), "nothing" ("nichts") or "never" ("nie"), for the latter have a clear excluding character. The last item "other" includes all other negation markers, which only appear rarely. However, comparing the distribution of negation markers between material products and services, no significant differences were found.

TABLE 1: MEAN TOKEN DISTANCE BETWEEN ATTRIBUTE AND NEGATION MARKER

Product Categories	Mean Token Distance
Services	1,7918
Material Products	1,9234
All	1,8576

In the following, the mean token distance between attribute and negation marker was examined (see TABLE 1). The results show that the distance between attribute and negation marker tends to be shorter within customer reviews relating to services than within customer reviews rating material products. However, the Kruskal-Wallis-Test

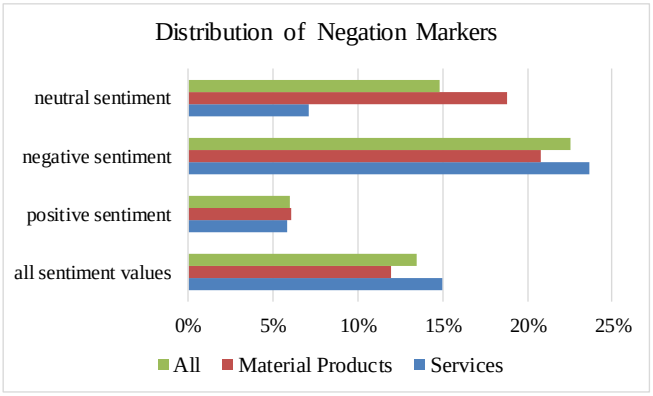

Figure 2: Distribution of negation markers

revealed that with a p-value > 0.05 the token distances are equally distributed. Therefore, the differences in the token distances are not significant.

Third, the frequency of negation marker use was examined between both categories. Figure 2 illustrates the percentages of negation markers used within the different product categories, as well as negation marker used in total. The use of negation markers is displayed with respect to the sentiment value of the sentence. On average, each sentence contained more than 1.7 marked statements respectively attributes evaluating (a part of) a product. Approximately 52% statements were assigned as positive, 41% as negative, and 7% as neutral. The values indicate a difference between the use of negation markers in customer reviews for material products and services. Regarding the negation markers used within all sentiment values, as well as negation markers used

 ISBN: 978-1-61208-600-2

within negative sentiment, customer reviews relating to services contain more negation markers. The results of the distributions also reveal that there seem to be differences in the use of negation markers between customer reviews about services and material products in general. However, a significance test was applied to test whether differences across sentiment values are significant.

For applying an appropriate significance test, we assume that the decisions, whether to use a negation marker associated with an attribute or not, are independent from one another and are the result of a Bernoulli distribution as the decision is either "yes" or "no". This includes a distributed random variable with success probability p, where p in {p_service, p_materialproduct} may or may not be different from customer reviews about services and those about material products. We test the null hypothesis, that p_service = p_materialproduct, with a binomial test using the statistical software R.

For the test, we consider the binomial distributed variable X_service that counts the number of negation markers used in the attributes in service reviews, and construct a confidence interval on a confidence level of 98% based on the annotated customer reviews. The confidence interval consists of the 0.01- and 0.99-quantiles of the distribution of X_service ~ B(p_service, n_materialproduct).

As a result, we observe, that p_materialproduct*n_ materialproduct is not located inside the confidence interval. We repeat the test for X_materialproduct ~ B(p_materialproduct, n_service) and find, that p_service*n_service is outside the obtained confidence interval.

Therefore, we postulate with a certainty of 98% that p_service != p_materialproduct. Concomitantly, the p-value <0.05 states that a significant difference between these two categories exists. As a consequence, we can confirm our hypothesis that customer reviews relating to services contain more negation markers than customer reviews rating material products.

Additionally, regarding the distributions of negations within neutral sentiment values (see Figure 2), there is a rather large difference in frequency recognizable. Although only 7% of the statements were marked as neutral, we conducted the significance test in the same way as for all sentiment values, but only for neutral sentiment. We receive a p-value <0.05 stating that there exist a significant difference. As an explanation, one might argue that customers judge products in a much less euphoric way than services. Thus, if the sentiment is neutral, for products, people might use formulations like "not so bad". In contrast, for services they preferably use e.g., "okay". However, as these are assumptions, it is necessary to examine these in more detail.

V. CONCLUSION

The results obtained in our study confirm our hypothesis about a difference in the use of negation markers in customer reviews rating services compared to customer reviews rating material products.

Human social interaction, as well as the personal commitment towards the person providing the service leads to a more polite writing style. When rating services customers rate the executing individuals and thus, are more moderate and polite in their judgement using negation markers instead of words containing a negative polarity value on a lexical basis. To prove this concept, in our future work we aim to investigate this assumption in more detail.

As our analysis also showed a difference in the use of negation markers in neutral sentiment, but with a reverse distribution, it would be interesting to examine these findings in detail as well. However, as neutral sentiment seems to be not that numerous in customer reviews, a special corpus for this issue needs to be compiled.

In addition, we examined the mean token distance between the attribute and the associated negation marker within the two product categories. In contrast to other features of language use, significant differences in the use of language could not be proven here. However, the mean token distance could still be a useful input variable for sentiment analysis of German customer reviews.

Generally, we strive to use our findings in the analysis of complaints from German customer reviews. In our future work, we aim to filter relevant information about products or services. If a company provides as well services as products, it would be very beneficial to identify automatically if customers speak about the product or about the service. Thus, the information could be allocated directly towards the right product type.

ACKNOWLEDGMENT

The support of the German National Science Foundation (DFG) through the funding of the research project "Entwicklung eines Sensors für ungerichtete Beschwerden aus Online Foren" (SCHM 1856/69-1) is gratefully acknowledged.

REFERENCES

[1] HolidayCheck Group, "Psychologie des Bewertens", 2016. [Online] Available from: https://www.holidaycheckgroup.com/wp-content/uploads/2016/07/TOMORROW-FOCUS-AG_Umfrage_Die-Psychologie_des_Bewertens_Pr%C3%A4sentation-1.pdf. [Retrieved: September, 2017].

[2] U. Farooq, H. Mansoor, A. Nongaillard, Y. Ouzrout, and M. A. Qadir, 2017, "Negation Handling in Sentiment Analysis at Sentence Level", JCP, 12(5), 2017, pp. 470-478.

[3] C. Diamantini, A. Mircoli, and D. Potena, "A Negation Handling Technique for Sentiment Analysis". In Collaboration Technologies and Systems (CTS), 2016 International Conference on, IEEE, October, 2016, pp. 188-195.

[4] A. Hogenboom, P. Van Iterson, B. Heerschop, F. Frasincar, and U. Kaymak,, "Determining negation scope and strength in sentiment analysis". In Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, October, 2011, pp. 2589-2594.

[5] A. Asmi and T. Ishaya, "Negation identification and calculation in sentiment analysis". In The Second International Conference on Advances in Information Mining and Management, 2012, pp. 1-7.

 ISBN: 978-1-61208-600-2

[6] I.G. Councill, R. McDonald, and L. Velikovich, “What's great and what's not: learning to classify the scope of negation for improved sentiment analysis”. In Proceedings of the workshop on negation and speculation in natural language processing. Association for Computational Linguistics, July, 2010, pp. 51-59.

[7] A. Kennedy and D. Inkpen, “Sentiment classification of movie reviews using contextual valence shifters”. Computational intelligence, 22(2), 2006, pp. 110-125.

[8] E. Ogiermann, 2009, “Politeness and in-directness across cultures: A comparison of English, German, Polish and Russian requests.” Journal of Politeness Research, 5, 2009, pp. 189-216.

[9] Stickel, G., “Untersuchungen zur Negation im heutigen Deutsch”, Vol. 1, Springer-Verlag, 2013.

[10] R. Giora, N. Balaban, O. Fein, and I. Alkabets, “Negation as positivity in disguise”. Figurative language comprehension: Social and cultural influences, 2005, pp. 233-258.

[11] E. Goffman, “Interaction ritual: Essays in face to face behavior.” AldineTransaction, 2005.

[12] S. Rogatcheva, 2004, “Politeness in English and German: a contrastive study” GRIN Verlag.

[13] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, “A survey on the role of negation in sentiment analysis”. In Proceedings of the workshop on negation and speculation in natural language processing, Association for Computational Linguistics, July, 2010, pp.60-68.

[14] G. Vinodhini and R. M. Chandrasekaran, “Sentiment analysis and opinion mining: a survey”. International Journal, 2(6), 2012, pp. 282-292.

[15] B. Pang and L. Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales”. In Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics, June, 2005, pp. 115-124.

[16] K. Boland, A. Wira-Alam, and R. Messerschmidt, “Creating an annotated corpus for sentiment analysis of german product reviews.” GESIS - Leibniz-Institut für Sozialwissenschaften. Mannheim, GESISTechnical Reports 2013/05, May, 2013.

[17] A. I. Metzmacher, V. Heinrichs, B. Falk, and R. H. Schmitt, “Customer Language Processing“. In Proceedings of the 12th International Workshop on Semantic and Social Media Adaptation and Personalization, IEEE, July, 2017, pp.32-33.

[18] J. Hedley “jsoup: Java HTML Parser”, 2009. [Online] Available from: https://jsoup.org/ [Retrieved: September, 2017].

[19] A. N. Rafferty and C. D. Manning, “Parsing three German treebanks: Lexicalized and unlexicalized baselines”. In Proceedings of the Workshop on Parsing German. Association for Computational Linguistics, June 2008, pp.40-46.

[20] M. L. McHugh, “Interrater reliability: the kappa statistic”, Biochemia medica, 22(3), 2012, pp. 276-282.

 ISBN: 978-1-61208-600-2