SEMAPRO 2023

The Seventeenth International Conference on Advances in Semantic Processing

ISBN: 978-1-68558-108-4

September 25 - 29, 2023

Porto, Portugal

SEMAPRO 2023 Editors

Karsten Böhmm, FH Kufstein Tirol Bildungs GmbH, Austria

SEMAPRO 2023

Forward

The Seventeenth International Conference on Advances in Semantic Processing (SEMAPRO 2023), held between September 25th and September 29th, 2023, continued a series of international events that were initiated considering 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 2023 constituted the stage for the state-of-the-art on the most recent advances.

We take here the opportunity to warmly thank all the members of the SEMAPRO 2023 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 who dedicated much of their time and effort to contribute to SEMAPRO 2023. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the SEMAPRO 2023 organizing committee for their help in handling the logistics of this event.

We hope that SEMAPRO 2023 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of semantic processing.

SEMAPRO 2023 Chairs

SEMAPRO 2023 Steering Committee

Sandra Lovrenčić, University of Zagreb, Croatia
Tim vor der Brück, Lucerne University of Applied Sciences and Arts, Switzerland
Michele Melchiori, Università degli Studi di Brescia, Italy
Wladyslaw Homenda, Warsaw University of Technology, Poland
Fabio Grandi, University of Bologna, Italy
Els Lefever, LT3 | Ghent University, Belgium
Sofia Athenikos, Twitter, USA

SEMAPRO 2023 Publicity Chairs

Laura Garcia, Universitat Politecnica de Valencia, Spain
Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain

SEMAPRO 2023
Committee

SEMAPRO 2023 Steering Committee

Sandra Lovrenčić, University of Zagreb, Croatia
Tim vor der Brück, Lucerne University of Applied Sciences and Arts, Switzerland
Michele Melchiori, Università degli Studi di Brescia, Italy
Wladyslaw Homenda, Warsaw University of Technology, Poland
Fabio Grandi, University of Bologna, Italy
Els Lefever, LT3 | Ghent University, Belgium
Sofia Athenikos, Twitter, USA

SEMAPRO 2023 Publicity Chairs

Laura Garcia, Universitat Politecnica de Valencia, Spain
Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain

SEMAPRO 2023 Technical Program Committee

Witold Abramowicz, Poznan University of Economics, Poland
Harry Agius, Brunel University London, UK
Jose María Alvarez Rodríguez, Carlos III University of Madrid, Spain
Abdel-Karim Al-Tamimi, Higher Colleges of Technology, UAE
Sofia Athenikos, Twitter, USA
Fernanda Baiao, PUC-Rio, Brazil
Arvind Bansal, Kent State University, USA
Giuseppe Berio, Université de Bretagne Sud | IRISA, France
Floris Bex, Utrecht University & University of Tilburg, Netherlands
Loris Bozzato, Fondazione Bruno Kessler, Trento, Italy
Zouhaier Brahmia, University of Sfax, Tunisia
Okan Bursa, Ege University, Turkey
Ozgu Can, Ege University, Turkey
Rodrigo Capobianco Guido, São Paulo State University (UNESP), Brazil
Damir Cavar, Indiana University, USA
Alberto Cetoli, QBE Europe, UK
David Chaves-Fraga, Universidad Politécnica de Madrid, Spain
Ioannis Chrysakis, FORTH-ICS, Greece / Ghent University, Belgium
Cláudio de Souza Baptista, Federal University of Campina Grande, Brazil
Christian Dirschl, Wolters Kluwer Deutschland GmbH, Germany
Milan Dojchinovski, InfAI | Leipzig University, Germany / Czech Technical University in Prague, Czech Republic
Julio Cesar Duarte, Instituto Militar de Engenharia, Rio de Janeiro, Brazil
Enrico Francesconi, IGSG - CNR, Italy
Rolf Fricke, Condat AG, Berlin, Germany
Panorea Gaitanou, Greek Ministry of Justice, Athens, Greece
Bilel Gargouri, MIRACL Laboratory | University of Sfax, Tunisia

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

Conceptual Semantic Evaluation Metric Using Taxonomy

İlknur Dönmez
Scientific and Technological Research Council of Türkiye
TÜBİTAK BİLGEM
Kocaeli, Turkey
email: ilknur.donmez@tubitak.gov.tr

Nur Bengisu Çam
Scientific and Technological Research Council of Türkiye
TÜBİTAK BİLGEM
Kocaeli, Turkey
email: nur.cam@tubitak.gov.tr

***Abstract*— The conceptual method is an important technique for calculating semantic similarity. In this study, we propose a taxonomy-based formula for calculating the conceptual similarity of sentences. The coefficients in the formula calculate how similar the noun words that make up the sentence ("verbs" and "adjectives" are also included) are to their most similar conjugates in the other sentence by considering the distance of these two words from their common ancestor and the position of the common ancestor in the ontology tree. We test our proposed metric in the English Semantic Textual Similarity (STS) benchmark dataset for semantic similarity. Although the labels of the dataset were not generated specifically for conceptual similarity, we were able to achieve 77 % accuracy in determining similar sentences using our proposed formula (which uses only noun types).**

Keywords-Similarity measures; word alignment; taxonomy; conceptual similarity, sentence similarity.

I. INTRODUCTION

In Artificial Intelligence (AI) and cognitive science, semantic similarity has become an established area of research to evaluate the strength of the semantic relationship between objects (such as words and documents). In recent years, a number of ontology-based semantic similarity metrics have been developed because they can mimic human cognitive functions. Among them, techniques based on the intrinsic information concepts have shown significant association with human evaluation [1].

According to Pirró and Euzenat [2], the scientific community divides the concepts of semantic measures into two main categories: Semantic Similarity (SS), which considers taxonomic relations such as "is-a" between two entities, and Semantic Relatedness (SR), which considers non-taxonomic relations between two entities (e.g., "cause-effect" and other associative relations such as fish lives-in water, where "lives-in" associates fish and water).

The semantic measure can be used in a variety of situations, e.g. in estimating similarity between documents [3], ontology-based text clustering [4][5], text summarization [6], entity disambiguation [7], developing recommender systems [8], semantic annotation [9], ontology merging [10], ontology segment matching [11], information retrieval [12], personalized support [13]-[15], and the graph editor similarity search problem [16], etc. Another important area is medical applications, which include automatic retrieval of patient records and medical documents [17]- [19].

The focus of this study is on semantic similarity, i.e., the "is-a" type relation between entities and ontologies is used as semantic evidence. The term "ontology" refers to any structure, such as a thesaurus, taxonomy, or other classification system, that formalizes knowledge without limiting its applicability.

Conceptual similarity comparison is an evaluation done by the human mind in order to understand semantics. The problem is to find the relationship between different concepts. In our study, after representing the sentence with its noun, verb, and adjective contents, we propose a semantic similarity metric to calculate the similarity distance between different sentences. We have made our codes publicly available on GitHub to ensure reproducibility and support future research [20].

The rest of the article is as follows: Section II contains related work on semantic similarity. Section III gives a general idea of the semantic representation of a sentence. Section IV introduces the novel similarity metric. Section V provides information about the dataset used and the evaluation results on this dataset. Section VI contains the final considerations of the metric and the results.

II. RELATED WORK

Semantic similarity of sentences has always been a popular research topic. In earlier times, methods evolved from looking at sentences word by word as a distinguishing feature to using grammatical rules to represent sentences [21]. After the creation of WordNet [22], a lexical database structured by semantic relations, ontology has been used by many researchers to compute the semantic similarity between words [23]- [28]. Jiang and Conrath measured the similarity of words by combining the taxonomy with the statistical information of the given corpus [24]. Seco et al. proposed to use WordNet for extracting the Information Content (IC) for computing the semantic similarity of words [25]. Yang and Powers proposed two different edge-based search approaches for similarity computation using WordNet [26]. Liu et al. computed the similarity between words by using the shortest part between words and the depth information from WordNet [27]. Similarly, Zhou et al. used the path length and IC value from WordNet [28]. They also compared their results with those of other authors, including Jiang and Conrath.

So far, we have mentioned the various approaches to calculating similarity between words. However, there are other studies that compute sentence similarities rather than

 ISBN: 978-1-68558-108-4

word similarities [29]- [33]. Sravanthi and Srinivasu analyzed the existing methods for computing sentence similarity and applied feature selection techniques for further investigation [29]. Selvarasa et al. used knowledge-based and corpus-based methods to measure sentence similarity in Tamil language [30]. Jeyaraj and Kasthurirathna proposed a multilayer semantic similarity network with the different number of layers and tested it on the SemEval [31] dataset [32]. Lee proposed a new approach for computing similarity between long sentences using WordNet [33].

III. Semantic Representation of A Sentence

It is still impossible to fully represent the semantic elements of a sentence or text in AI. Geoffrey Leech suggests seven types of meaning, namely "conceptual, connotative, social, affective, reflective, collocative, and thematic", in his book "Semantics: The Study of Meaning" [34]. Semantic features depend on the meaning of words, word relationships, their position in the whole context (contextual features), their emphasis, references to the physical world such as color, time, geological location, natural laws, rhetoric, and even the understanding of the reader [35], [36].

In semantics, the concept is about “What is the text or sentence about and what does it refer to?". These are also the first questions we ask when we try to understand a text. Once we know the concepts, we can move on to the important relations, attributes, orders, and references. Before determining the similarity score in our study, we determined the nouns, verbs, and adjectives in the sentences and made a list for each one, as shown in Table I.

TABLE I. Preprocessing

Sentence 1: A woman is dancing and singing with other women. **Sentence 2:** A girl is dancing and singing in the rain.			
	Noun	**Verb**	**Adjective**
S1	Woman	Dancing, singing	Other
S2	Girl, rain	Dancing, singing	-

IV. Proposed Metric

Having presented each sentence as in the example in Table I, how can we determine whether or not the words in sentences are similar? It is not hard to see similarities if the words are not the same. We know from our daily lives that the human brain can understand the relationship between subordinate and superordinate words (hyponyms and hypernyms).

If we find the similarity value for each pair of words, we can average them to calculate the similarity between sentences. Equation (1) shows our proposed formula to calculate the similarity between sentences. We compare each word in one sentence with the words in the other sentence, and the most similar pairs of words are included in the calculation of the average.

$$sim_i = max_j \begin{cases} 1, & n_{1,i} = n_{2,j} \\ \frac{h_{cp,i,j}}{(h_{cp,i,j} - h_{1,i} + 1)(h_{cp,i,j} - h_{2,j} + 1)}, & n_{1,i} \neq n_{2,j} \end{cases} \quad (1)$$

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

$$sim_{i\_s} = max_j \begin{cases} 1, & n_{1,i} = n_{2,j} \\ \frac{h_{cp,i,j}}{(h_{cp,i,j} - h_{1,i} + 1)(h_{cp,i,j} - h_{2,j} + 1)(h_{max})}, & n_{1,i} \neq n_{2,j} \end{cases} \quad (3)$$

As can be seen in Figure 1, the starting node is the root element of the ontological tree; when we talk about the WordNet, the root word is “entity”. $n_{1,i}$ is the i^{th} word in the first sentence, and $n_{2,j}$ is the j^{th} word in the second sentence. To calculate the similarity between $n_{1,i}$ and $n_{2,j}$, if $n_{1,i}$ and $n_{2,j}$ are the same, their similarity value is 1. For each i^{th} element of sentence 1, we find the similarity value for all j^{th} words in sentence 2, and the maximum similarity is considered.

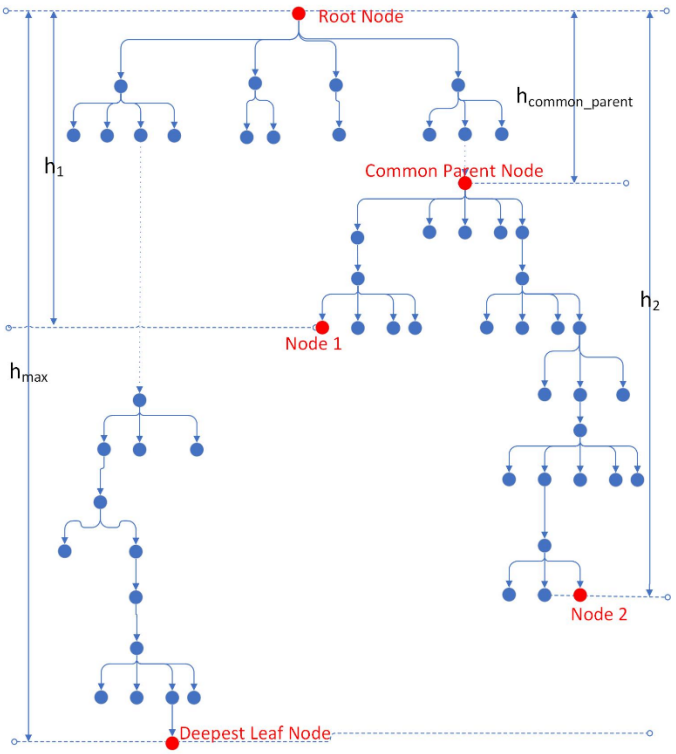

Figure 1. Nodes and their heights.

If the nodes are not equal, the distance is correlated with the height difference of the nodes to the common parent (n_{cp}). If both children are closer to the common parent, it means that the concepts of the children's nodes are also closer and similar. When the distance to the common parent is small, the similarity is high.

 ISBN: 978-1-68558-108-4

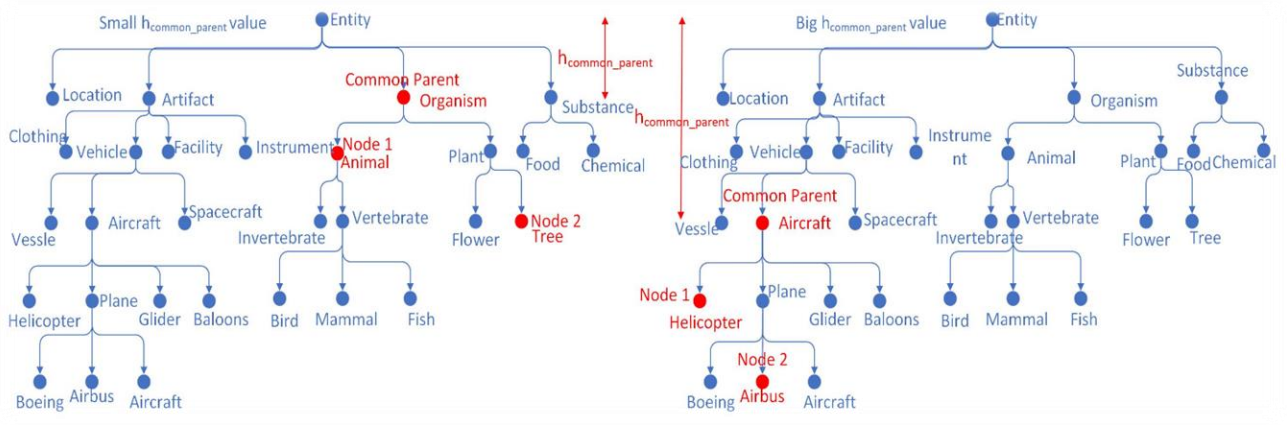

Figure 2. The depth of the common parent effect on similarity.

In the formulas given, $h_{cp,i,j}$ is the depth of the common parent node for the i^{th} word of sentence 1 and the j^{th} word of sentence 2. When we fix the common parent-child distance, the children are dissimilar when the common parent is closer to the root element. When a common parent is closer to the leaf node, its children are more similar, as shown in Figure 2. Consider the entity node. It initially has two children, one living and one non-living. However, when we go to the deepest nodes in the ontology, the two child nodes of "motorcycle" become more similar. They could be, for example, "motor scooters" and "mopeds".

Since the minimum similarity value of (1) is equal to zero and the maximum similarity value of (1) is equal to h_{max}, using min-max normalization yields the normalized similarity value as in (3). Figure 2 shows the depth of the common node effect in a simplified version of WordNet. In the left block, the depth of the common parent node is large compared to the root node. The parent node is closer to the leaf nodes than the nodes in the right block. Thus, in the left block, the children are more similar than in the second block, even though the depth difference between children and common parent is the same for the two examples in the right and left blocks. We can say that the depth of the common parent is inversely proportional to the similarity of the word.

The similarity of two different sentences is calculated using (4). n is the total number of features and similarity is the distance for each feature. If some words do not have a pair, the average similarity value decreases when divided by the number of words. We also consider the synonym-sets because a word may have more than one meaning and we take the average to decrease the error.

$$sim = \frac{1}{n}\sum_{i=1}^{n}(sim_i) \quad (4)$$

Our algorithm is as follows:

- Step 1: For each pair of sentences in the dataset, remove the stop-words and the punctuations.
- Step 2: For each sentence in a pair, extract the Part-Of-Speech (POS) tags of each word.
- Step 3: Create a combination of the words in the pair according to their POS tags, then calculate the similarity score of the word pairs, using (3).
- Step 4: From the previous step, we have many similarity scores for a word. Accept the maximum similarity score.

V. Dataset & Evaluation

We used the train split of the English STS benchmark dataset [37] to evaluate our proposal for computing semantic similarity. This dataset is a collection of data given in SemEval tasks between 2012 and 2017. It contains sentence pairs and their similarity scores. There are 5749 sentence pairs in the train split. The given similarity scores range from 0 to 5, where 0 means that the pairs have no similarity, and 5 means that the pairs are equally similar. These scores are annotated by human judges. To increase readability, we normalized the similarity scores using the min-max normalization function of scikit-learn [38].

We computed the pair similarity scores of the train split of the English STS benchmark dataset. First, the similarities are computed by considering only the nouns in the sentences. Second, the similarities are computed by considering both nouns and verbs in the sentences. For calculating the depth of the nodes in the taxonomy, we used WordNet. WordNet is a large electronic lexical database for English that proposes a hierarchical structure of concepts, where lower elements inherit information from their parents [22].

The similarities between the nouns in the sentences and the similarities between the verbs are averaged at the end to calculate the final similarity of the pairs. At last, the similarities are calculated by considering the nouns, verbs, and adjectives in the sentences. Again, the similarities between the nouns of the sentences, the similarities between the verbs, and the similarities between the adjectives are averaged at the end to calculate the final similarity of the pairs.

 ISBN: 978-1-68558-108-4

TABLE II. NORMALIZED PAIR SIMILLARITY SCORES EXAMPLES

Pairs	Sentences	STS Similarity Score	Noun Only Similarity Score	Noun + Verb Similarity Score	Noun + Verb + Adjective Similarity Score
1	A woman is dancing and singing with other women.	0.60	0.73	0.86	0.56
	A girl is dancing and singing in the rain.				
2	Two men are packing suitcases into the trunk of a car.	0.88	1.00	0.75	0.50
	The men are putting suitcases into the car's trunk.				
3	The woman picked up the kangaroo.	0.75	1.00	0.50	0.33
	A woman picks up a baby kangaroo.				
4	Two foxes are eating from a plate on a brick patio.	0.56	0.51	0.75	0.50
	Foxes are eating from a plate.				
5	Two zebras are playing.	0.85	1.00	0.50	0.34
	Zebras are socializing.				
6	A group of people dance on a hill.	0.64	0.67	0.33	0.22
	A group of people are dancing.				
7	A car is moving through a road.	0.80	1.00	0.50	0.33
	A car is driving down the road.				
8	The man is shooting an automatic rifle.	0.76	0.58	0.79	0.52
	A man is shooting a gun.				
9	A woman is cutting up a chicken.	0.55	0.54	0.27	0.18
	A woman is slicing meat.				
10	Butter is being put into a bowl.	0.85	1.00	0.50	0.33
	A man cutting butter into a mixing bowl.				

If any of the sentences of the pairs do not contain adjectives, then the similarity between the adjectives is zero. Therefore, the similarity score in such a case is drastically lower when we consider the similarity of the adjectives. In Table II, we have given ten sentence pairs with their normalized STS similarity scores as well as the similarity scores we calculated.

As can be seen from Table II, finding the nouns in the sentences and calculating the similarity of these nouns according to the proposed formula yields a meaningful similarity criterion.

Our similarity criterion is based on conceptual knowledge. As Lawrence W. B. Barsalou said, “The human conceptual system contains people's knowledge of the world. Conceptual knowledge in the conceptual system supports a variety of basic cognitive operations, including categorization, inference, and the representation of propositions.” [1].

To get an idea of how the proposed method works with the dataset, we chose different threshold values. When the similarity threshold for a set of normalized similarity values is set to 0.50, any value greater than or equal to 0.50 is considered similar.

For these sets, we examined what percentage of these sentence pairs had a similarity of 0.50 or greater. Similarly, we chose the threshold value of 0.25 to understand the percentage of pairs that are not strongly similar and not strongly dissimilar. In Table IV, we have given the percentage results for different word selections and threshold settings.

During the evaluation, we performed the following tests for nouns, noun+verb, and noun+verb+adjective. We used the train split of the English STS benchmark dataset, which has normalized similarities between 1.00–0.75, 0.75–0.50, and 1.00–0.50, and checked what percentage of pairs the proposed method finds in this range to see if our proposed method also labels these pairs similarly. Again, we used the same data set, which has normalized similarities between 0.50 – 0.25, 0.25 – 0.00, and 0.50 – 0.00, meaning that the pairs are not similar. We checked what percentage of pairs the proposed method finds in this range to see if our proposed method also names them similarly.

Our Pearson correlation results can be found in Table III. Here, we measured the correlation between the normalized similarity values of the data set and the normalized similarity values we calculated.

 ISBN: 978-1-68558-108-4

TABLE III. CORRELATION RESULTS

Task Name	Pearson Correlation Score
Noun Only	0.51
Noun + Verb	0.47
Noun + Verb + Adjective	0.48

TABLE IV. EVALUATION

	Results of the Proposed Method	
Noun Only	Percentage of pairs in the range of 1.00– 0.50	77.09%
	Percentage of pairs in the range of 1.00 – 0.75	72.42%
	Percentage of pairs in the range of 0.75 – 0.50	82.05%
	Percentage of pairs in the range of 0.50 – 0.00	56.85%
	Percentage of pairs in the range of 0.50 – 0.25	42.38%
	Percentage of pairs in the range of 0.25 – 0.00	69.45%
Noun + Verb	Percentage of pairs in the range of 1.00 – 0.50	56.58%
	Percentage of pairs in the range of 1.00 – 0.75	50.74%
	Percentage of Pairs in the range of 0.75 – 0.50	62.77%
	Percentage of pairs in the range of 0.50 – 0.00	71.94%
	Percentage of pairs in the range of 0.50 – 0.25	59.37%
	Percentage of pairs in the range of 0.25 – 0.00	82.88%
Noun + Verb + Adjective	Percentage of pairs in the range of 1.00 – 0.50	46.35%
	Percentage of pairs in the range of 1.00 – 0.75	19.47%
	Percentage of pairs in the range of 0.75 – 0.50	74.88%
	Percentage of pairs in the range of 0.50 – 0.00	81.99%
	Percentage of pairs in the range of 0.50 – 0.25	54.76%
	Percentage of pairs in the range of 0.25 – 0.00	95.71%

We ran our experiments on an Intel(R) Core(TM) i7-9750H CPU. The entire experiment took about 48 seconds.

VI. CONCLUSION

In this study, we have proposed a formula for calculating the conceptual similarity of sentences. In the formula, the coefficients calculate how similar the noun words that make up the sentence ("verbs" and "adjectives" are also included) are to their most similar conjugates in the other sentence by looking at the distance of these two words to their common ancestor and the location of the common ancestor in the ontology tree. If the compared words are close to their common ancestor, they are more likely to be similar. The other important parameter is the depth of the common ancestor in the ontology tree. If the common ancestor is far from the root, the similarity of the compared words increases according to its position closer to the root node.

Since we are interested in the conceptual similarities, even if WordNet also has a taxonomic structure of adjectives that indicate the attribute of the nouns (concepts), they are not the actual concepts [39]. The inclusion of the similarity contribution between verbs or adjectives in our study negatively affected the results. This may be understandable if we consider similarity as a conceptual method.

In our upcoming research, we want to use a dataset where the conceptual similarity of sentences is scored by humans. To decide on a 5-level similarity scale (high similarity, low similarity, different, completely different, and no idea), participants are asked to use crowd-sourcing methods. How meaningful the results are determined by comparing the proposed similarity calculation with the human markers. We expect our method to give better results on the human tagged datasets, since our proposed method simulates the human mind to find the conceptual relationship.

REFERENCES

[1] L. W. Barsalou, "The human conceptual system," The Cambridge handbook of psycholinguistics, pp. 239-258, 2012.

[2] G. Pirró and D. Talia, "UFOme: An ontology mapping system with strategy prediction capabilities," Data & Knowledge Engineering, vol. 69(5), pp. 444-471, 2010.

[3] F. Benedetti, D. Beneventano, S. Bergamaschi, and G. Simonini, "Computing inter-document similarity with context semantic analysis," Information Systems, vol. 80, pp. 136-147, 2019.

[4] J. Nasir, I. Varlamis, A. Karim, and G. Tsatsaronis, "Semantic smoothing for text clustering," Knowledge-Based Systems, vol. 54, pp. 216-229, 2013.

[5] W. Song, J. Liang, and S. Park, "Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering," Information Sciences, vol. 273, pp. 156-170, 2014.

[6] S. Kumar and K. Bhatia, "Semantic similarity and text summarization based novelty detection," SN Applied Sciences, vol. 2, pp. 1-15, 2020.

[7] A. Vretinaris, C. Lei, V. Efthymiou, X. Qin, and F. Özcan, "Medical entity disambiguation using graph neural networks," In Proceedings of the 2021 International Conference on Management of Data, pp. 2310-2318, 2021.

[8] V. Demertzi and K. Demertzis, "A hybrid adaptive educational eLearning project based on ontologies matching and recommendation system," arXiv preprint, arXiv:2007.14771, 2020.

[9] A. Chikkamannur, "Semantic Annotation of IoT Resource with ontology orchestration," In 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC- IEEE) pp. 1-7, 2020.

[10] S. Mhammedi, H. El Massari, and N. Gherabi, "Composition of large modular ontologies based on structure," In Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C'21, pp. 144-154, 2022.

[11] A. Belhadi, Y. Djenouri, G. Srivastava, and J. Lin, "Fast and Accurate Framework for Ontology Matching in Web of Things," ACM Transactions on Asian and Low-Resource Language Information Processing, 2023.

 ISBN: 978-1-68558-108-4

[12] G. Yang et al., “CCGIR: Information retrieval-based code comment generation method for smart contracts,” Knowledge-Based Systems, 237, 107858, 2022.

[13] M. Sreenivasan, S. Dhar, and A. Chacko, “PCPS: Personalized Care through Patient Similarity,” In 2022 IEEE Region 10 Symposium (TENSYMP-IEEE), pp. 1-6, 2022.

[14] Y. Liu and M. Ijaz, “Personalized auxiliary information presentation system for mobile network based on multimodal information,” Mobile Networks and Applications, pp. 1-11 2022.

[15] F. Liu and S. Li, “Research on personalized user-centered product improvement based on sentiment mining of online reviews and competitor analysis,” Home, http://www.researchsquare.com/article/rs-1829215/v1 (Accessed Aug. 19, 2023).

[16] S. Babalou, A. Algergawy, and B. KönigRies, “SimBio: Adopting Particle Swarm Optimization for ontology-based biomedical term similarity assessment,” Data & Knowledge Engineering, 102137, 2023.

[17] M. Landolsi, L. Hlaoua, and L. Ben Romdhane, “Information extraction from electronic medical documents: state of the art and future research directions,” Knowledge and Information Systems, vol. 65(2), pp 463-516, 2023.

[18] B. Yang et al., “Classification of Medical Image Notes for Image Labeling by Using MinBERT,” Tsinghua Science and Technology, vol. 28(4), pp. 613-627, 2023.

[19] D. Tian, M. Li, Y. Shen, and S. Han, “Intelligent mining of safety hazard information from construction documents using semantic similarity and information entropy,” Engineering Applications of Artificial Intelligence, vol. 119, 105742, 2023.

[20] N. B. Cam, Sementic Similarity. Github. https://github.com/bengisucam/semanticSimilarity . (accessed Aug. 19, 2023).

[21] R. P. Honeck, “Semantic similarity between sentences,” Journal of Psycholinguistic Research, vol. 2, pp. 137-151, 1973.

[22] G. A. Millar, “WordNet: A Lexical Database for English.,” In Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

[23] M. Sussna, “Word sense disambiguation for free-text indexing using a massive semantic network,” In Proceedings of the second international conference on Information and knowledge management, pp. 67-74, 1993.

[24] J. J. Jiang and D. W. Conrath “Semantic similarity based on corpus statistics and lexical taxonomy,” arXiv preprint cmp-lg/9709008, 1997.

[25] N. Seco, T. Veale, and J. Hayes, “An intrinsic information content metric for semantic similarity in WordNet,” In Ecai vol. 16, p. 1089, 2004.

[26] D. Yang and D. M. Powers, “Measuring semantic similarity in the taxonomy of WordNet,” Australian Computer Society, 2005.

[27] X. Y. Liu, Y. M. Zhou and R. S. Zheng, “Measuring semantic similarity in WordNet,” In 2007 international conference on machine learning and cybernetics, vol. 6, pp. 3431-3435, 2007.

[28] Z. Zhou, Y. Wnag, and J. Gu, “New model of semantic similarity measuring in wordnet,” In 2008 3rd International Conference on Intelligent System and Knowledge Engineering, vol. 1, pp. 256-261, 2008.

[29] P. Sravanthi and B. Srinivasu, “Semantic similarity between sentences,” International Research Journal of Engineering and Technology (IRJET), vol. 4, no. 1, pp. 156-161, 2017.

[30] A. Selvarasa, N. Thirunavukkarasu, N. Rajendran, C. Yogalingam, S. Ranathunga, and G. Dias, “Short Tamil sentence similarity calculation using knowledge-based and corpus-based similarity measures,” In 2017 Moratuwa Engineering Research Conference, pp. 443-448, 2017.

[31] SemEval 2016 Dataset: https://altqcri/semeval2016/task2. (Accessed Aug. 19, 2023)

[32] M. N. Jeyaraj and D. Kasthurirathna, “Mnet-SIM: A multi-layered semantic similarity network to evaluate sentence similarity,” International Journal of Engineering Trends and Technology, vol. 69, no. 7, pp. 181–189, 2021. doi:10.14445/22315381/ijett-v69i7p225

[33] M. C. Lee. “A novel sentence similarity measure for semantic-based expert systems,” Expert Systems with Applications, vol. 38, no. 5, pp. 6392-6399, 2011.

[34] N. Love, “Translational semantics: A discussion of the second edition of Geoffrey Leech's Semantics: The Study of Meaning,” Stellenbosch Papers in Linguistics, vol. 11, pp. 115-136, 1983.

[35] D. Geeraerts, “Theories of lexical semantics,” OUP Oxford, 2009.

[36] T. A. Van Dijk, “Society and discourse: How social contexts influence text and talk,” Cambridge University Press, 2009.

[37] Cer, Daniel et al., "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," arXiv preprint arXiv:1708.00055, 2017.

[38] Pedregosa, Fabian et al., "Scikit-learn: Machine learning in Python," The Journal of machine learning research 12, pp. 2825-2830, 2011.

[39] D. Gross and K. J. Miller, “Adjectives in WordNet,” International journal of lexicography, vol. 3, no. 4, pp. 265-277, 1990.

 ISBN: 978-1-68558-108-4

DYNAMO: Dynamic Ontology Extension for Augmenting Chatbot Intelligence through BabelNet

Amalia Georgoudi
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
ageorgoudi@iti.gr

Georgios Meditskos
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
gmeditsk@csd.auth.gr

Thanassis Mavropoulos
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
mavrathan@iti.gr

Stefanos Vrochidis
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
stefanos@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
ikom@iti.gr

***Abstract*—Dynamic ontology extension, a real-time ontology extension process, facilitates continuous learning and adaptation to new knowledge and evolving domains. The ability to dynamically add concepts, relationships, and axioms allows intelligent agents and knowledge-based systems to stay up-to-date and responsive. This paper presents a novel approach to dynamically extend the ontology of a chatbot's Knowledge Base by leveraging BabelNet, a multilingual encyclopedic dictionary and semantic network. Using BabelNet's semantic relations, such as hyperonyms, hyponyms, and holonyms, we focus on enriching the ontology with user profile information, enabling knowledge inference and personalized interactions. Through repeated interactions, the chatbot increases its level of intelligence, inferring new knowledge and asking targeted questions to users, resulting in effective interactions and increased user satisfaction.**

***Index Terms*—Dynamic Ontology Extension; Ontology; Chatbot.**

I. Introduction

Ontologies play a vital role in improving the efficiency and effectiveness of retrieving information and representing knowledge online. Ontologies serve as organized structures for capturing knowledge and establishing a shared language to describe concepts, relationships, and properties within specific domains. Through explicit definitions of terms and their relationships, ontologies enable seamless integration, searching, and reasoning of information. They facilitate the creation of machine-readable and machine-understandable representations, fostering interoperability and data exchange across diverse systems and domains.

A chatbot is an Artificial Intelligence (AI) program specifically created to mimic human conversation and engage with users through interfaces that use text or voice. By incorporating ontologies into a Chatbot's Knowledge Base (KB), the chatbot would have access to a rich semantic network of concepts, relationships, and properties within a specific domain. Ontologies define the vocabulary and the meaning of terms, enabling the chatbot to understand user input and generate relevant and accurate responses.

Dynamic ontology extension is the process of extending or updating an existing ontology in real-time, depending on new information or shifting requirements. In order to enable the ontology to change and adapt to new knowledge or evolving domains, it entails dynamically adding new concepts, relationships, attributes, or axioms to it. Intelligent agents, semantic web applications, and knowledge-based systems all heavily depend on dynamic ontology extension in order to continuously learn, adapt, and incorporate new information.

Through the dynamic extension of the chatbot's KB ontology, based on user inputs, the chatbot gradually enhances its intelligence with each interaction. With the ability to expand its ontology in real-time, the chatbot becomes increasingly proficient at understanding user needs and providing tailored responses, leading to a more natural and effective user experience.

In this paper, we present our method for the dynamic ontology extension of a chatbot's KB ontology by utilizing BabelNet [11]. BabelNet is a public multilingual encyclopedic dictionary and a semantic network, in which every entity is connected with other entities through semantic relations, such as hypernyms, hyponyms and holonyms. We utilized those relations and we created a pattern, in order to automatically extend the ontology with more information regarding the users profiles, enabling inferencing knowledge, enhancing their understanding, and drawing conclusions. The goal is to enrich the situational awareness of the chatbot every time it interacts with the user, reducing the time required for the interaction and increasing user satisfaction. The proposed method extends the schema of the ontology with new classes and properties and also populates the ontology by adding new instances and forming semantic relationships and links between them.

The remainder of the paper is organized as follows. Section II presents the related work. In Section III the proposed

 ISBN: 978-1-68558-108-4

method is introduced, while in Section IV a simulation example is presented. In Section V the evaluation procedure is presented. Section VI concludes and gives directions for future work.

II. Background and Related Work

The following subsections provide a brief overview of Knowledge Representation frameworks, Ontology-based Chatbots and existing Dynamic Ontology Extension approaches.

A. Knowledge Representation

An ontology is a structured framework for organizing information that provides a formal and explicit specification of a commonly recognized formulation of a field of interest. The representation of knowledge as a set of concepts, relationships, and properties is part of this formatting. A straightforward knowledge representation language known as the Resource Description Framework (RDF) [1] intends to standardize metadata and usage descriptions for Web-based resources. A group of subject-predicate-object triples serve as the fundamental building unit of RDF. To display richer and more complicated information about objects, groups of things, and relationships between them, the Web Ontology Language (OWL) [2], a collection of knowledge representation languages, was developed. OWL now extends previous Web standards for describing knowledge and is the official World Wide Web Consortium (W3C) recommendation for creating and sharing ontologies. An OWL ontology consists of: classes, properties and restrictions. Classes are the main element of an ontology and are used to describe a field's concepts. Properties are used to describe feature attributes, while restrictions are determining properties' confinement. Furthermore, an ontology has instances of classes and relationships between those.

B. Ontology-based Chatbots

Utilizing ontologies is a potential strategy that enables chatbots and conversational agents to comprehend and produce contextually relevant responses. In this section, many studies that present ontology-based chatbots and conversational agents are examined, highlighting their contributions and outlining their ramifications.

The SynchroBot [3] is a dialog system that can be connected to reliable and adaptable KBs and KGs for information extraction and that can use NLP tools to analyze user questions and NLG techniques to deliver appropriate answers. A previous ontology-based chatbot called OntBot used Hallili's method as its foundation. OntBot [4] uses a suitable mapping approach to convert ontologies and other knowledge into relational databases with a set of mapping rules.

The recent bibliography includes numerous proposed methods for introducing KGs to QA and AI chatbots in different domains, such as E-commerce [5], museums [6], healthcare systems [7]. A working model of Ontology based chatbot that handles queries from users for an E-commerce website, is proposed in [5]. This chatbot helps the user by mapping relationships of the various entities required by the user, thus providing detailed and accurate information there by overcoming the drawbacks of traditional chatbots. The author in [6] summarizes recent research on Knowledge Graph (KG)-based AI chatbot design and development for museums. The suggested MuBot approach gave museums the chance to develop chatbots for their visitors. The use of KGs for chatbot implementation raises issues with translating natural language dialogues

In [7], the suggested conversation agent, is built on an ontology-based knowledge model that enables flexible reasoning-driven dialogue planning as opposed to the use of predetermined dialogue scripts. Another comparable strategy was used to the healthcare industry with the intention of creating a framework that may help patients by giving them access to an AI chatbot with good conversational abilities and a substantial KB [9] [8]. MediBot [10] is another ontology-based chatbot created to facilitate access to information on drugs and their risks easily and directly to Portuguese speakers.

C. Dynamic Ontology Extension Frameworks

Ontology evolution is a subfield of ontology change which refers to the problem of transforming an ontology in response to a certain need [12]. Specifically, ontology evolution consists of transforming an ontology or incorporating new information in an existing ontology, in a way that it satisfies the users and describes the knowledge domain, while maintaining its consistency.

In survey [13], the authors outline the process of ontology learning and categorizes ontology learning techniques into three classes: linguistics [14], statistical [19] [16], and logical [19] that are based on reasoning rules. It examines the advantages and disadvantages of ontology evaluation techniques and it explores the applications and significance of ontology learning in different industries.

Linguistic-based approaches for ontology extension utilize NLP tasks, such as Named Entity Recognition, Part-Of-Speech Tagging and Dependency Parsing and they have been proposed for many different languages and domains. For instance, there are approaches for Spanish legal texts [15], French [17] and Chinese [18] documents.

A statistical and logical system, known as CRCTOL is proposed in [19]. CRCTOL is a system designed to automate the extraction of ontologies from domain-specific documents. The proposed system, CRCTOL, utilizes a comprehensive text parsing technique and incorporates a combination of statistical and lexico-syntactic methods and includes a a statistical, a word sense disambiguation and a rule-based algorithm.

In addition, some proposed methods utilize Deep Learning models, such as [20] [22]. In [20], a novel approach to automatically expand ontologies, specifically focusing on the ChEBI ontology, a well-established reference in the field of chemistry within life sciences, is presented. To achieve this, the authors utilized a deep learning model called ChemBERTa [21], which is built upon the Transformer architecture. The model was trained using the leaf node structures found in the ChEBI ontology along with their corresponding classes.

 ISBN: 978-1-68558-108-4

In [22], a bi-LSTM-based word extraction model based on character embedding is presented to extract the terms from a phrase for automatic ontology population.

Our approach distinguishes itself by leveraging publicly available resources, such as BabelNet and Wikidata. These resources play a vital role in shaping the classes and their hierarchical relationships within the ontology, employing semantic relations like hyponyms and holonyms. Furthermore, our approach utilizes dependency parsing, a process to grammatically analyze sentences, to establish direct and indirect data and object properties. Our methodology ensures that multilingual information is preserved, thereby enabling multilingual interactions. By incorporating these distinct elements, our approach presents a novel and robust framework for ontology development and facilitates enhanced linguistic and semantic interoperability.

III. Dynamic Ontology Extension Approach

This section provides a high level overview of the architectural design of the dynamic ontology extension approach. Our approach extends the schema of the ontology with new classes and properties and also populates the ontology by adding new instances and forming semantic relationships and links between them.

The method is built upon a language analysis task that specifically targets named entity recognition. The solution has been implemented utilizing the Python programming language, harnessing a range of libraries, including Owlready2, Babelnet, SPARQLWrapper, among others. Through this solution, the ontology is augmented by incorporating entities identified as persons, organizations, or geopolitical locations. This enriched information plays a pivotal role in enabling a chatbot to draw accurate conclusions and gain a deeper understanding of the users' background, thereby facilitating more effective interactions.

The approach encompasses four distinct phases, each serving a specific purpose in the overall ontology extension process. These phases are: the Disambiguation, the Formulation of Classes' Hierarchy, the Population of the Ontology and finally the Generation of Data and Object Properties.

A. Disambiguation

When an entity, mentioned in the speaker's utterances, is classified as a Named Entity, and contains a Babelnet id, the service will be enabled. The entity tag will constitute the classes, while the entities will constitute the instances. Three different cases are supported:

1) The class does not exist in the ontology.
2) The class already exists, but the specific instance does not exit.
3) Both the class and the specific instance already exist in the ontology.

The first case is when the class does not exist in the ontology. Our service creates the class, the necessary properties, and adds the new instance. In case that the class already exists, our service creates only the new instance of this class. Otherwise, if both the class and the specific instance exist, our service will not extend further by adding duplicates into the ontology.

Moreover, an additional disambiguation phase is conducted for entities categorized as geopolitical locations or organizations, utilizing BabelNet relations. Within the disambiguation phase for geopolitical location entities, the primary objective is to ascertain whether the entity represents a country or a city. In the case of a city, a further check is performed to determine if it serves as a capital. Additionally, for every entity that is tagged as an organization, the goal is to understand the type of the organization. For instance an organization may be a university, a hospital, a bank, etc. This disambiguation process relies on BabelNet synsets and ISA relationships to accomplish the aforementioned tasks.

B. Formulation of Classes' Hierarchy

During the second phase, the algorithm proceeds with the creation of classes and subclasses, following a hierarchy derived from BabelNet relationships. This process leverages the information acquired during the disambiguation phase, wherein hypernyms, hyponyms, and holonyms associated with each entity tag are utilized to effectively structure the classes within the ontology.

A hypernym relation refers to a hierarchical relationship where one entity is more general or broader in meaning than another entity. It signifies that the first entity encompasses or includes the second entity within its scope. Holonyms in BabelNet denote the connection between a complete entity and its component parts. A holonym represents the entirety of the entity, whereas its parts are identified as meronyms.

By utilizing these semantic relationships, the algorithm ensures a coherent and organized representation of the entities within the ontology, facilitating better categorization and classification. Figure 1 illustrates an example of the classes that were created form the Geopolitical Entities (GPE) instances "Greece" and "Sweden", by iteratively exploiting BabelNet relationships.

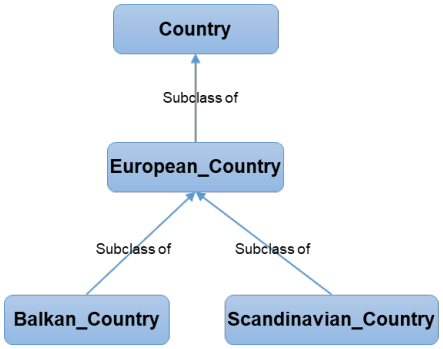

Fig. 1. Hierarchy of classes example.

C. Population of the Ontology

Following the successful construction of the ontology's classes, the third phase involves the population of the ontology. Within this phase, the entities will undergo classification as

instances of specific classes, determined by their assigned named entity tags. Subsequently, the formation of links and connections between distinct entities will take place.

To avoid duplicate elements within the ontology, the service incorporates a mechanism that checks for the existence of classes and properties before creating new ones. For instance, if an entity is associated with the ”person” tag, a new class named ”person” will only be generated if it does not already exist in the ontology. Similarly, each time an entity is tagged as ”person,” a new instance will be added to the corresponding class, ensuring that instances are not duplicated. This approach maintains the integrity and coherence of the ontology by preventing redundant class and instance creation.

D. Generation of Data and Object Properties

For the data properties, we retrieve important information from the BabelNet in order to form the triples. First of all, each entity is associated with a data property that stores its unique BabelNet ID. Furthermore, we store the information in multiple languages, allowing the chatbot to engage in multilingual conversations effectively. Additionally, we enhance the data by incorporating supplementary details from Wikidata, utilizing SPARQL queries specifically tailored for each geopolitical location. This comprehensive approach ensures that the ontology encompasses diverse and enriched data, enabling the chatbot to provide accurate and contextually relevant information during interactions. Table I shows some examples of data properties.

TABLE I
DATA PROPERTIES

Data Property	***Description***	***Range***
official_languages	The official languages of a country are the languages that are recognized and used by the government for official business and communication	string
babelnet_id	The unique babelnet if of every entity	string
has_value	Gives the name in different languages	string
geo_lat	Geographical coordinates: Latitude	float
geo_long	Geographical coordinates: Longitude	float

Relations between entities in the ontology are also created. Also, inverse properties and transitive properties are created between two entities. Table II represents some object properties that can be created between different entities, such as cities and countries.

TABLE II
OBJECT PROPERTIES

Object Property	***Description***	***Domain***	***Range***
is_city_of	A city is in a country	City	Country
has_city	A country has a city	Country	City
is_in_city	An organization is in a city	Organization	City
Has_organization	A city has an organization	City	Organization

Furthermore, the ontology is extended with direct and indirect relations between the speaker and the entities, depending on what the user said. Figure 2 represents the created direct and indirect relationships between the speaker and the entities, based on the user's utterance *”I live in Thessaloniki with my friend Maria.”*. In order to achieve this, we parse relations in the dependency tree to connect indirect entities.

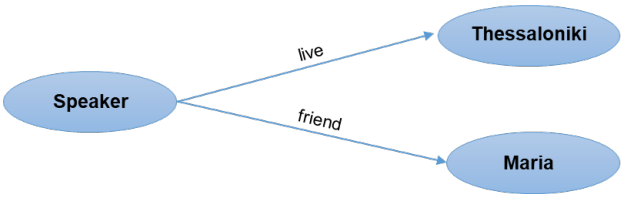

Fig. 2. Direct and indirect relations with the speaker.

IV. EXAMPLE USE CASES

Below, we will describe two simulation examples of the dynamic ontology extension task. In the first example, the initial ontology is empty. Assuming the following user response:

User: *”I would like to ask how to go to the National and Kapodistrian University of Athens”*

In this example, the module generates four object properties, five datatype properties, eight new classes, and four instances.

Regarding the second example, the input ontology is the one previously created. Assuming the user utterance is as follows:

User: *“I live in Thessaloniki with my friend Maria.”*

In this case, two new instances, “Thessaloniki” and “Maria”, are added to the “Person” and “City” classes, respectively. Additionally, two new object properties, named “live” and “friend”, are added to establish relationships between the speaker and the new instances.

In the ontology extension process, when an entity is labeled as “PERSON,” a new class named “Person” is created if it does not already exist. Subsequently, for each entity labeled as “PERSON,” a new instance is added to the corresponding “Person” class. Similarly, if an entity is classified as “GPE”, after the disambiguation phase, a new class named “City” or “Country” is created if it is not already present. If a class named “City” is created and the entity's status as a capital is verified using ISA relations from BabelNet, a subclass named “Capital” is established under the “City” class. Then, to further expand the ontology dynamically, a new class named ”Country” is introduced along with subclasses indicating the continent to which the country belongs.

In the case where the language analysis tool identifies an entity as an organization, a new class named “Organization” is generated. Moreover, a subclass is created to indicate the organization's type, which is retrieved from hypernym relations from BabelNet. Finally, the new instance, representing the organization, is added. Furthermore, the city and country names associated with the organization are included in the ontology, and the relevant classes are created accordingly, if do not already exist.

 ISBN: 978-1-68558-108-4

V. EVALUATION

A. Ontology Evaluation

1) Debugging: In order to ensure thorough evaluation, we executed the debugger for every extended ontology. This comprehensive approach allowed us to assess the output ontology and its extensions with greater scrutiny. The employment of Protégé as the initial development tool, coupled with the utilization of the 'Pellet' reasoner and the 'Debugger' plug-in, enabled us to search for any potential faults in the ontologies. Remarkably, the debugging process confirmed the absence of faults during the validation process for all extended ontologies.

2) Ontology Metrics: The structure of the output ontology, that was presented in Section IV was evaluated using OntoMetrics, an online framework designed to validate ontologies based on established metrics. The findings obtained from the analysis conducted by OntoMetrics are provided in Table III. The metrics presented in Table III encompass both simple metrics, such as the count of classes, axioms, and objects, as well as schema metrics. The simple metrics, categorized as Base Metrics, provide insights into the number of ontology elements. On the other hand, schema metrics focus on the design aspects of the ontology, reflecting its richness, width, depth, and inheritance.

TABLE III
ONTOLOGY METRICS GENERATED BY ONTOMETRICS.

Category	Metric	Value
Basic	Axioms	108
Basic	Class Count	10
Basic	Object Property Count	7
Basic	Data Property Count	6
Basic	Individual Count	6
Basic	Description Logic Expressivity	ALI(D)
Schema	Attribute richness	0.6
Schema	Inheritance richness	0.9
Schema	Relationship richness	0.4375
Schema	Axiom/class ratio	10.8
Schema	Inverse relations ratio	0.285714
Schema	Class/relation ratio	0.625

B. Application-based evaluation

It is important to acknowledge that the proposed method achieves around 98% accuracy rate in expanding the ontology for entities categorized as locations, persons, or organizations. This achievement highlights the efficacy and robustness of the proposed approach, establishing its potential as a valuable tool for ontology expansion and enhancement in domains where locations, persons, and organizations play a significant role.

Furthermore, the proposed approach successfully fulfills its objective of enhancing the agent's intelligence and significantly reducing the time required for information retrieval. This efficiency is achieved through the generation of new knowledge autonomously, eliminating the need to ask certain questions. Consequently, this reduction in interaction time not only enhances user satisfaction but also reinforces the perception of the agent's advanced intelligence. Moreover, the approach enables seamless multilingual interactions by comprehending entity names in various languages. This feature proves particularly beneficial for individuals who may lack fluency in verbal communication, thus promoting inclusivity and accessibility in human-agent interactions. By encompassing these capabilities, the proposed approach demonstrates its potential to improve user experience and facilitate effective communication across diverse linguistic backgrounds. Leveraging the geographical coordinates, the chatbot possesses the capability to determine the migrant's location accurately. Consequently, it can provide tailored recommendations of nearby places, organizations, and public sectors based on the user's specific inquiries.

In the rest of this Section, we will present a compilation of competency questions alongside the corresponding SPARQL queries and their exceptionally satisfactory outcomes.

CQ1: Given a specific city, determine the corresponding country and identify the official language spoken within that locality.

The following SPARQL query returns the country that has a city named "Thessaloniki" and the official languages of the specific country.

```
SELECT ?country ?language
WHERE {
  ?country rdf:type ex:country.
  ?country ex:has_city ex:Thessaloniki.
  ?country ex:official_languages ?language
     .}
```

Listing 1. SPARQL Query CQ1

The output (Table IV) encompasses the country names along with their corresponding official languages.

TABLE IV
RESULTS CQ1

a/a	Country	official_languages
1	Greece	"Greek"
2	Greece	"Demotic_Greek"

CQ2: Given a specific organization, determine the type of the organization in order to understand a user's background and work experience.

The following SPARQL query returns the type of a given organization.

```
SELECT ?class
WHERE {
  ex:UBS a ?class .}
```

Listing 2. SPARQL Query CQ2

Table V displays the outcomes obtained from CQ2.

CQ3: Given a specific city name in German, return the name of the city in French.

The next SPARQL query yields the French translation of the city name "Thessaloniki," despite it being stored in German.

 ISBN: 978-1-68558-108-4

TABLE V
RESULTS CQ2

a/a	Classes
1	owl:Thing
2	Organization
3	Bank

```
SELECT ?frenchName
WHERE {
  ?city rdf:type ex:City;
       ex:has_value "Thessaloniki"@de;
       ex:has_value ?frenchName.

  FILTER (LANG(?frenchName) = "fr")}
```

Listing 3. SPARQL Query CQ3

Table VI illustrates the output of the SPARQL query CQ3, which is the French translation of the city "Thessaloniki".

TABLE VI
RESULTS CQ3

a/a	French Name
1	"Thessalonique"@fr

VI. CONCLUSION

In summary, this research paper presented a dynamic ontology extension approach that leverages relationships from BabelNet. The approach was implemented within an ontology-based chatbot system. The results demonstrate that with each user interaction, the chatbot continually enhances its intelligence by integrating new information into the ontology. This progressive intelligence augmentation leads to significant reductions in interaction time and heightened user satisfaction, as users perceive the chatbot's increasing inteligence. Additionally, the proposed method enables the chatbot to engage in multilingual conversations, effectively bridging language barriers and accommodating users with diverse backgrounds.

Overall, this research highlights the potential of the dynamic ontology extension approach in creating more intelligent and adaptable chatbot systems. Future investigations could explore further advancements in utilizing BabelNet relationships and extend the multilingual capabilities to enhance user experience in a broader range of linguistic contexts.

ACKNOWLEDGEMENT

This work has received funding by the European Commission as part of its H2020 Programme under the contract number 870930-IA (WELCOME), grant agreement No. GA101017558 (ALAMEDA), and grant agreement No. GA101049294 (LEAGUE).

REFERENCES

[1] S. Decker et al., "The semantic web: The roles of XML and RDF," IEEE Internet Computing, 4, 63-73, 2000.
[2] L. Ma et al., "Towards a complete OWL ontology benchmark," 3rd European Semantic Web Conference, ESWC 2006 Budva, Montenegro, June 11-14, 2006, pp. 125-139, 2006.
[3] A. Hallili, "Toward an ontology-based chatbot endowed with natural language processing and generation," 26th European Summer School In Logic, Language and Information, 2014.
[4] H. Al-Zubaide and A. Issa, "Ontbot: Ontology based chatbot," International Symposium On Innovations In Information And Communications Technology, pp. 7-12, 2011.
[5] A. Vegesna, P. Jain, and D. Porwal, "Ontology based chatbot (for e-commerce website)," International Journal Of Computer Applications, 179, 51-55, 2018.
[6] S. Varitimiadis, K. Kotis, D. Spiliotopoulos, C. Vassilakis, and D. Margaris, "Talking triples to museum chatbots," Culture And Computing: 8th International Conference, Held As Part Of The 22nd HCI International Conference, 2020, pp. 281-299, 2020.
[7] L. Wanner et al., "Kristina: A knowledge-based virtual conversation agent," Advances In Practical Applications Of Cyber-Physical Multi-Agent Systems: The PAAMS Collection: 15th International Conference, Porto, Portugal, 2017, pp. 284-295, 2017.
[8] E. Kamateri et al., "Knowledge-based intelligence and strategy learning for personalised virtual assistance in the healthcare domain," Proceedings Of Semantic Technologies For Healthcare And Accessibility Applications (SyMpATHY), 2019.
[9] G. Meditskos et al., "Towards an ontology-driven adaptive dialogue framework, Proceedings of The 1st International Workshop On Multimedia Analysis And Retrieval For Multimodal Interaction, pp. 15-20, 2016.
[10] C. Avila et al., "MediBot: an ontology based chatbot for Portuguese speakers drug's users," International Conference On Enterprise Information Systems, 1 pp. 25-36, 2019.
[11] R. Navigli and S. Ponzetto, "BabelNet: Building a very large multilingual semantic network," Proceedings Of The 48th Annual Meeting Of The Association For Computational Linguistics, pp. 216-225, 2010.
[12] G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, and G. Antoniou, "Ontology change: classification and survey," The Knowledge Engineering Review, 23, 117-152, 2008.
[13] M. Asim, M. Wasim, M. Khan, W. Mahmood, and H. Abbasi, "A survey of ontology learning techniques and applications, Database, 2018.
[14] P. Cimiano and J. Völker, "A framework for ontology learning and data-driven change discovery," Proceedings Of The 10th International Conference On Applications Of Natural Language To Information Systems (NLDB), 3513, pp. 227-238, 2005.
[15] J. Völker, S. Langa, and Y. Sure, "Supporting the Construction of Spanish Legal Ontologies with Text2Onto," Springer, 2008
[16] E. Drymonas, K. Zervanou, and E. Petrakis, "Unsupervised ontology acquisition from plain texts: the OntoGain system," 15th International Conference on Applications of Natural Language To Information Systems, NLDB 2010, June 23-25, 2010, pp. 277-287, 2010.
[17] M. Hajji, M. Qbadou, and K. Mansouri, "An adaptation of Text2Onto for supporting the French language," International Journal Of Electrical and Computer Engineering (2088-8708), 10, 2020.
[18] X. Jing, G. Yan-Qiang, S. Ai-Ju, and X. Jian-Liang, "Chinese ontology learning technology based on Text2Onto," 2012 7th International Conference On System Of Systems Engineering (SoSE).
[19] X. Jiang and A. Tan, "CRCTOL: A semantic-based domain ontology learning system," Journal Of The American Society For Information Science And Technology, vol. 61, pp. 150-168, 2010.
[20] A. Memariani, M. Glauer, F. Neuhaus, T. Mossakowski, and J. Hastings, "Automated and explainable ontology extension based on deep learning: A case study in the chemical domain," ArXiv Preprint ArXiv:2109.09202, 2021.
[21] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," ArXiv Preprint ArXiv:2010.09885, 2020.
[22] M. Su, C. Wu, and P. Shih, "Automatic ontology population using deep learning for triple extraction," 2019 Asia-Pacific Signal And Information Processing Association Annual Summit And Conference (APSIPA ASC), pp. 262-267, 2019.

 ISBN: 978-1-68558-108-4

Semantically Augmented Documents for Use in Higher Education Institutions

Analyzing the Current State in the Digital Transformation of HEI

Karsten Böhm
FH Kufstein Tirol University of Applied Sciences
Kufstein, Austria
email: karsten.boehm@fh-kufstein.ac.at

Abstract—Higher Education Institutions (HEI) are part of the Digital Transformation of our societies and a higher level of digitization and automatization is already establishing within the sector. The Information and Communication Technology (ICT) landscapes, the data models and the processes are highly individual to the HEI and interoperability and linking of data is a challenge and often creates new digital format discontinuities. This contribution analyses the support level of recent semantic models, namely the European Learning Model and the Educational Verifiable Credentials Model for the application in an Austrian university with respect to graduation documents. The results show the application potential of the new models and relates it to current practices in representing data of academic programs in a meaningful way. An implementation demonstrates use-cases with an immediate effect for HEI and focuses on attractive and lightweight User Experience (UX) to ensure user adoption.

Keywords—Digital Transformation; HEI; Linked Data; European Learning Model; Verifiable Credentials Model.

I. INTRODUCTION

Higher Education Institutions (HEI) are offering education programs with a more or less planned learning journey to reach specified qualification objectives within a qualification framework (e.g., the European Qualification Framework [1]). As a result of the Bologna process of aligning the national education systems in Europe, the education system became more transparent and interoperable – on a national, but also on an European level. While being a positive development for students, this also meant an increasing number of stakeholders for the HEI. They need to be integrated in the internal processes and information systems, e.g., when Recognizing Prior Learning (RPL) for a study program [2] or to include student mobility into a course program. Together with the development of a digital transformation in many organizations, this results for HEI in a higher level of digitization and automatization that is already establishing within the sector. However, the ICT landscapes, the data models and the processes are highly individual to the HEI and interoperability and linking of data is a challenge. Often, this situation creates new digital format discontinuities and requires additional efforts for the organizations, staff and students of a HEI.

At the other hand, there are a number of interesting developments at the European level that try to harmonize and digitize the information exchange within and among HEI. Among the most notable are the new version of the Education Learning Model (ELM) [3] and the W3C standard on the Verifiable Credential Model (VCM) [4] that recently had been extended with a version for the educational sector, the Educational Verifiable Credential Model (EVC) [5]. Both models address the need to formalize and harmonize the unstructured information and document sources that limit the automated processing of relevant documents that maybe already digital in format but very diverse with respect to the semantic representation.

Common ways to document the achievements of academic programs are the two document categories Diploma Supplement (DS) and Transcripts of Records (ToR). While a DS describes the general program aspects and the individual properties of the student, the ToR contains information about the subjects taken and the grade for each subject [6], [7]. Both documents aim at describing the specific aspects of the program and the achieved result of the student in great detail and are mandatory to be generated by Austrian universities for all graduates. However, currently those documents are using a predefined structure, the content is only text based. DS and ToR need to be validated and durable, even when the course program or even the organization is not operational anymore. Therefore, it is important that the document is self-contained and independent from an organization or a technical system.

A ToR is already specifying qualifications, but only as textual data. For automated processing and the support of all stakeholders in a HEI (student, lecturers, managers) it would be useful to have that in a machine-readable way, as the information is often generated from ICT-systems anyway.

This contribution builds on the concept of educational Pre-built Information Spaces (PreBIS-ED) [8] and is researching the current state of the art concerning certification documents with a focus on the situation in Austria and a specific university as an illustrating example. It aligns the features of the aforementioned semantic models with the existing specification documents required in Austria and elsewhere in Europe. By doing so, it assesses the potential of the semantic models to be used in everyday processes within HEI and it identifies the gaps that still exist.

It contributes to the development of the digital excellence of HEI by addressing two important needs: 1) For HEI, the ability to process existing qualifications, e.g., from a first study cycle with the reference to RPL, Recognition of Prior Learning and 2) for companies it would be beneficial to have an opportunity to match qualifications against their job profiles, e.g., when doing a Job Task Analysis (JTA), [9].

While addressing a very specific and practical application area, PreBIS-ED also tries to find new answers for the more general research question on how (existing) semantic models can be put in operational practice by different stakeholders

 ISBN: 978-1-68558-108-4

that are not knowledge engineering experts. The approach focuses on two main aspects 1) putting the model to use with a clear benefit for the user (increased transparency and consistency in this case) and 2) creating an attractive and lightweight UX for areas that are manually crafted anyway (curricular structures and competences in this case).

The rest of the paper is structured as follows: In the Section 2, the current situation in an Austrian university with respect to the ToR and the DS documents is introduced – a situation that is typical for the Austrian sector and very common for universities all across Europe. Afterwards the predefined structure of the ToR and DS documents is matched against the semantic models ELM and EVC to determine the usefulness for those document categories. In Section 4, the implementation of a technical solution is outlined that brings the semantic models together with the existing document renderings. The paper concludes in Section 5 with an outlook on the next steps within these research activities.

II. The Case of an Austrian University

The HEI of the author is already using DS and ToR documents that are supplied automatically to each student that graduates from a bachelor's program or master's program. The documents are generated by the internal campus management system as PDF documents and are digitally signed using the electronic signature supplied for all public administration services in Austria [10]. The organization follows the rules set up by the European Commission and is granted the Diploma Supplement Label and the ECTS Label (ECTS – European Credit Transfer System) for adhering to those standards [11].

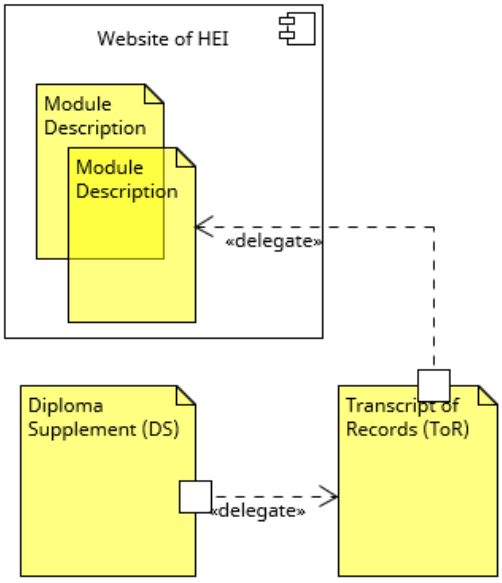

Figure 2. Relation of the different specification documents which are delivered to the graduating student (DS and ToR) and contain references to the module descriptions which are available online at the HEI website.

The documents contain more detailed information about the formal properties of the degree and also overview information about the qualification and learning objectives. In the case of this HEI the respective Section in the Diploma Supplement (Section 4.3) refers to the ToR document that contains more detailed information about the subjects in the degree program and the individual student results.

For competences and learning outcomes, it refers to the website of the HEI that contains detailed information about the curriculum and the learning outcomes in the general module description. Figure 1 illustrates the relation between the different information sources.

Currently, the DS/ToR does not contain any information about specific learning outcomes (but refers the user to the website of the university as an external information source) and is not semantically enriched. Therefore, the documents provide the potential for an extension with semantic information about specifically acquired competences during the program in order to support the role of interfacing between different HEI or the Human Resources departments of companies.

The information on the website is representing a static version of the program as such as it describes the learning outcomes in more general way; its implementation in the different semesters can vary in (within the defined scope).

The current situation with respect to the documentation of competences can be summarized in the following way:

(a) Information about the learning outcomes is stored distributed in three different locations two of which are delivered as documents to the students at the time of graduation and the third is stored online at the HEI website.
(b) The information is human readable only and needs interpretation and research depending on the level of detail needed.
(c) Due to its distributed nature, it is not self-contained and requires the availability of additional online-services to be used.
(d) As the documentation references the module description of a program and not the yearly made syllabus, it might not cover specific details, such as the used technology or software system used in the execution of a lecture in a certain semester, thus lacking some level of detail.

With respect to a more seamless integration of the information into other systems or organizations, it would be beneficial, if those properties could be improved by helping users of this documentation to have a more complete and easier to use data collection at their disposal. As the documentation is generated by an internal campus management system that is being developed by the HEI itself, there is the possibility to develop an improved solution to address this need.

III. Matching the ToR/DS to the Semantic Models

In order to estimate the potential of the concepts of the semantic models for improving the expressiveness of the ToR and DS document categories, the individual properties of each characteristic will be matched against the concepts from the

 ISBN: 978-1-68558-108-4

semantic models. It should be noted that the attributes of a ToR document and a DS document are defined by the European Commission and augmented by National Authorities [12], but the values are mostly free-text form, individual for each HEI and also subject to change. This makes the information less interoperable.

Using semantic models like the ELM and the EVC provide a scaffolding that could be helpful to create more consistent and interoperable representations for the values of the DS/ToR documents. In order to approve the suitability of this approach, the fields of DS and ToR documents are mapped to the semantic models in the next Sections.

A. Analysis of the different features in a DS

To add semantic information to a DS document, both the ELM and the EVC are good candidate models to encode the information in the current documents with an explicit semantic specification. Table I below provides a mapping of the fields defined for a DS document and available in the current DS of the author's HEI with classes (sometimes with properties) from the semantic models ELM and EVC.

TABLE I. ATTRIBUTES OF A DS DOCUMENT (LEFT COLUMN) AND APPROPRIATE SEMANTIC CLASSES AND PROPERTIES (RIGHT COLUMN). ABBREVIATIONS: EVC – EDUCATIONAL VERIFIABLE CREDENTIAL, VC – VERIFIABLE CREDENTIAL, ELM – EUROPEAN LEARNING MODEL

1. INFORMATION IDENTIFYING THE HOLDER OF THE QUALIFICATION	
1.1 Last name(s)	VC: holder
1.2 First name(s)	VC: holder
1.3 Date of Birth	VC: holder
1.4 Student identification number	VC: holder
2. INFORMATION IDENTIFYING THE QUALIFICATION	
2.1 Name of qualification, title conferred	EVC: Credential
2.2 Main field(s) of study for the qualification	EVC: Credential Subject
2.3 Name and status of awarding institution	EVC: Issuer
2.5 Course languages	ELM: language, default language (properties)
3. INFORMATION ON THE LEVEL AND DURATION OF THE QUALIFICATION	
3.1 Level of the qualification	ELM: QF level (property)
3.2 Official duration of program in credits and/or years	ELM: Credit Points
3.3 Access requirement(s)	
4. INFORMATION ON THE PROGRAM COMPLETED AND THE RESULTS OBTAINED	
4.1 Mode of study	
4.2 Program learning outcomes	ELM: Learning Outcome
4.3 Program details, individual credits gained and grades/marks obtained	ELM: Learning Achievement
4.4 Grading scheme, grade translation and grade distribution guidance	ELM: Grading Scheme
4.5 Overall classification of the qualification	ELM: Qualification, Qualification Reference
5. INFORMATION ON THE FUNCTION OF THE QUALIFICATION	
5.1 Access to further study	ELM: Learning Entitlement Specification
5.2 Access to a regulated profession (if applicable)	ELM: Learning Entitlement, Learning Entitlement Specification
6. ADDITIONAL INFORMATION	
6.1 Additional information	(Arbitrary textual information)
6.2 Further information sources	(Arbitrary textual information and URLs to Web Resources)
CERTIFICATION OF THE SUPPLEMENT	
7.1 Date	EVC: Issuance Date
7.2 Signature	EVC: Proof (Digital)
7.3 Capacity	EVC: Proof
7.4 Official stamp or seal	EVC: Credential Proof

It is shown that all the concepts of a DS can be captured in the new semantic models and thus provide a good starting point for an interoperable implementation. The study also shows that EVC focusses more on the administrative information of the DS whereas ELM focusses on the expressiveness of the education related aspects like qualifications and learning outcomes.

For the use case of understanding the (overall) qualification of a DS holder, Sections 3 and 4 of a DS are the most interesting parts and the modelling of ELM is more important for that field of application. Since ELM is developed by the European Commission and the DS document is a mandatory document for Austrian (and probably also for most European) universities, there is a good chance that the semantic encoding of the already defined fields can lead to interoperable and machine-readable specifications. Further evidence for this assumption was collected with a workshop series with members from different European Universities in the CloudEarthI-project [13] held in 2022 and 2023.

Currently, to the best knowledge of the author, no implementations of semantic information on the DS documents have yet been implemented by Austrian universities. This might be due to the fact that version 3 of ELM is just about to be released in its final version, according to [3]. The contribution can be thought as an initial activity to bring the emerging semantic models into operational effectiveness. Once this is achieved with application cases like augmenting the DS and ToR document, additional benefits can be addressed like automated status reports on inconsistencies with learning outcomes or credit point achievement, missing topics and matching qualifications. Machine readable semantic representations will also help to automate processes of validation and verification in a reliable way as it does not depend on ambiguous text representations. This will contribute to accelerate the Digital Transformation of the HEI domain.

B. Analysis of the different features in a ToR

While the DS only holds the overall qualification information, the ToR should document qualifications and learning outcomes in a more detailed level. Usually, a ToR document contains the information represented in Table II, below. This is also the case in the author's HEI. The usual

 ISBN: 978-1-68558-108-4

tabular presentation is ordered by the chronological order of the study program, e.g., into semesters in the case of the author's HEI. The items in Section 2 of Table II will occur for each subject grouped by the semesters in which they are positioned. In addition to the values for the lectures there are also derived information that are calculated per semester, e.g., the amount of credits (usually 30 per semester) and the average grade over all subjects studied in one semester.

TABLE II. ATTRIBUTES OF A TOR DOCUMENT (LEFT COLUMN) AND APPROPRIATE SEMANTIC CLASSES AND PROPERTIES (RIGHT COLUMN). ABBREVIATIONS: EVC – EDUCATIONAL VERIFIABLE CREDENTIAL, VC – VERIFIABLE CREDENTIAL, ELM – EUROPEAN LEARNING MODEL

1. INFORMATION IDENTIFYING THE HOLDER OF THE QUALIFICATION	
Last name(s)	VC: holder
First name(s)	VC: holder
Date of Birth	VC: holder
Student identification number	VC: holder
2. INFORMATION ON THE DIFFERENT LECTURES (GROUPED INTO SEMESTERS)	
Course Title	ELM: Identifier
Course Code	ELM: Identifier
Language of lecture	ELM: language (property)
Contact Hours of lecture	ELM: contact hours (property)
Credit point of lecture	ELM: credit received (property) ELM: volume of learning or workload (property)
Credit points per semester	ELM: credit points (property)
Grade (value and per cent)	(calculated value, information only)
Average grade per semester	(calculated value, information only)
CERTIFICATION OF THE SUPPLEMENT	
Date	EVC: Issuance Date
(Digital) Signature	EVC: Proof
Capacity	EVC: Proof

It is interesting that there are expected similarities in the Sections of the holder information and the certification between DS and ToR, but it is even more noteworthy that the ToR does not contain information about detailed qualifications and learning outcomes in a more specific way.

The only hint on the content of the respective lectures are the names of the lectures that might suggest the content and possible learning opportunities. Relevant information should be found elsewhere, in the case of the author's HEI at the website of the course program. The link between the ToR and the relevant content on the website is made by the name or course code of the lecture only.

This analysis shows that there is a high potential of embedding semantic information into the ToR to make competence information more visible at the level of individual lectures. ELM provides a number of useful classes to model and link this information together such as the classes "Learning Achievement" and the class "Learning Achievement Specification" with the properties "learning outcome" and "learning outcome summary".

Linking could be achieved by using the property "content URL" to refer to the website content with deep linking. This, however could also create the problem of broken links, if the structure of a website changes or if the online resource is not available. In the use case of ToR and DS document the information also need to remain static in the sense that it needs to reflect the information that was current at the time of document creation. Changes in the linked resources could even introduce semantic errors, if the updates content does not fit anymore. It would be better to embed the information into the document directly and use the linked information only as a secondary or supplemental resource. Since the curricular information on the website of the author's HEI are also generated from the same data source as the DS and ToR documents, this is technically feasible to implement.

C. Summary on the potential of semantic models

After mapping the semantic models to the ToR and DS documents it becomes clear that the current application case for the EVC is mostly targeted at modelling the outcomes (certificates) of the program, while the ELM in its new version is more versatile to model important concepts on several levels in the HEI. As illustrated in Figure 2, it can be used at the academic process level to support the operational processes and – most prominently – support the certification process by providing documents that are easier to process by interfacing stakeholders (e.g., other HEI or companies that are

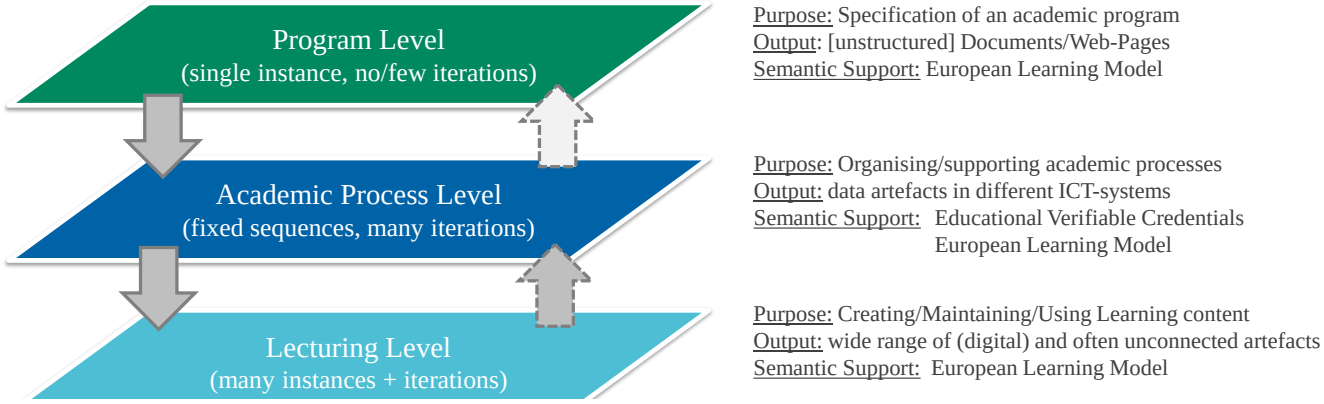

Figure 2. Overview on the different levels in the planning and the execution of academic programs with their semantic support levels by existing semantic models

 ISBN: 978-1-68558-108-4

hiring the graduates). Apart from this middle level, the semantic model could also support the design process (top-level) for aligning the learning objectives among the different modules and towards the qualification profile. Likewise, the operational activities of the lecturers can be supported by using the ELM in tools that help them to create course structure and materials that are linked to each other using learning outcomes.

IV. IMPLEMENTATION

The implementation of a DS and ToR documents with extended semantic information could be provided by different means and does also differ depending on the document format. The most important formats for this application case being HTML for the embedding in web-oriented applications (e.g., the HEI website, a web-based e-learning-system) and PDF for issuing the documents for students and other stakeholders. The use of metadata, microformats and embedded JSON-LD data could be a suitable approach for HTML, as this information is also recognized by search engines as structured markup and can be used to improve the search results based on facts and not only on text snippets [14] [15]. It seems that JSON-LD will be the preferred structured data format in the future by Google, according to the Search-EngineJournal [16], which is compatible with ELM and EVC, since both provide JSON-LD representations.

Simple meta-data formats (as key-value pairs) are also available for PDF, but another promising way is the use of the attachment feature of the PDF-format [17]. There are a number of different ways to relate existing files to a PDF document such as associating or referencing files, but for the purpose of this application the embedding of files as a file stream into the container PDF is the most suitable way, as it provides (a) the property of self-containment and (b) can be identified, used and even exported by the user. Using a suitable library such as PDFLib [18] makes it possible to augment the already generated document with semantic information, e.g., using the JSON-LD format [19] that is also used with ELM. This way the existing generation process does not need to be changed and the augmentation with the semantic information can be added as an additional step in the process.

Furthermore, the use of the attachment feature also has the benefit that the users can identify the augmented data, if the PDF software used, is supporting it (e.g., Adobe Acrobat or the PDF-viewer in Mozilla Firefox). This way, it can be downloaded and used wherever this might be useful. Since the implementation uses a standard functionality of the PDF specification it is agnostic to the tools used. Figure 3 shows an application in the Adobe Acrobat Reader with embedded semantic information as a JSON-LD-file in form of an attachment. The use of this approach would be the most feasible one for the DS and ToR documents as they are self-contained and could be easily represented in the document without any additional IT-system; they are interoperable and preserve the self-containment feature of the document.

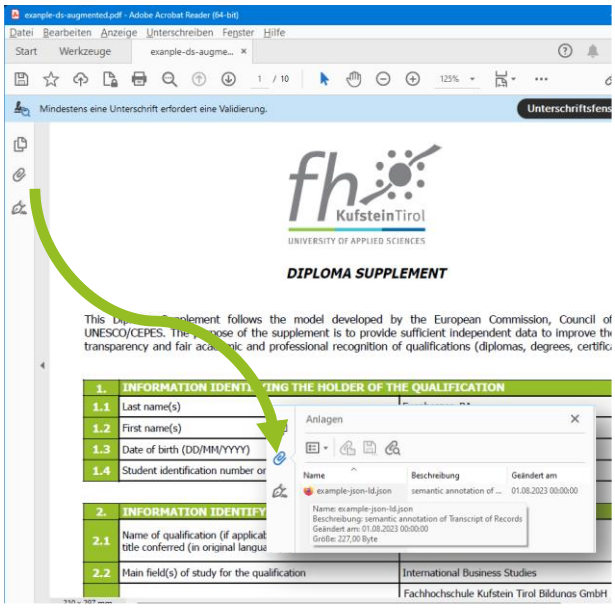

Figure 3. Screenshot of Adobe Acrobat as an example PDF application showing the embedded semantic information as JSON-LD data in a containerized representation.

While the representation of DS and ToR provide a good interfacing to external entities based on the concept of verifiable credentials (EVC) and a detailed domain model for the education domain (ELM), they are not suitable for a more detailed insight into the structure of the learning process that links the qualification objectives (c.f. Section 4.2 in the ToR document) to the individual learning paths that are usually fulfilled by lectures taken by the students. This currently hidden path could contribute to the transparency of the learning process, both during and after the student has completed her study. The ELM could be used to model the path and thus support the upper and the lower level of an academic program specification, as depicted in Figure 2.

It should be noted that the curricular structures are already crafted manually and usually encoded into a textual representation. This approach should help users to model the curricular structures by using an existing model (ELM) with the assistance of an attractive and lightweight user interface that hides the complexity of the semantic modelling by focusing on the instance level and only the needed concepts and properties for the application case. Hence the notion of Pre-Build Information Models, which remove the task of modelling from the user of the implementation. Currently the focus on the implementation is therefore on the visualization and use of the semantic models and less on automated concept extraction from existing textual curricular descriptions using text-mining technologies.

In order to visualize and use such a path though the curriculum the implementation provides two views that should be easy to understand and to navigate. It builds on the concept of Hierarchical Competence Maps (HCM), which are described in [20]. Figure 4 shows a detailed view on a single modular element (e.g., a module, a lecture or a teaching unit) with a few competences and their level of expertise,

 ISBN: 978-1-68558-108-4

visualizing the hierarchy with the metaphor of stacked cards, employing the concept of an HCM and using ELM as a data model for the semantic representation.

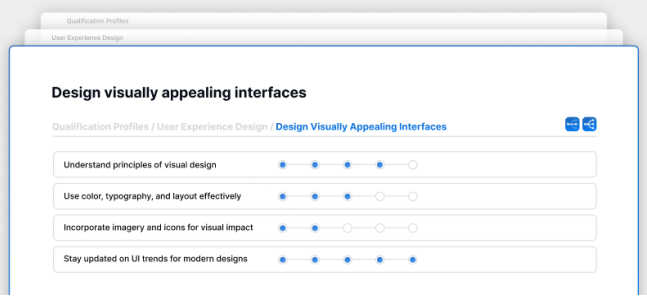

Figure 4. Screenshot of the detailed view of a modular element of an education program, using HCM and ELM.

The intention here is to focus on the immanent competences of the current element providing a detailed view on the learning outcome. As this view is lacking the overview of the connected modular elements this is provided by another view, shown in Figure 5 with the thread from the top-layer to the bottom layer highlighted. This overview should visualize the connection between different elements, helping the user to understand the relation and contributions between different parts of a curriculum. Two variations of the overview are implemented, one that is showing all linked data elements and another one that is emphasizing on a specific chain of competences from the (top-level) learning objective of the program to the learning goals of an individual lecture (at the most detailed level), using connected ELM concepts.

The user can switch between those two visualizations at any time and the implementation is carried out as a web-application that can be embedded in other web-applications, such as the website of a HEI or the e-Learning-system as needed.

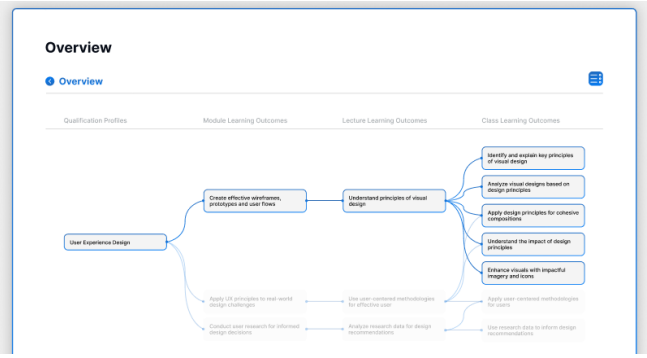

Figure 5. Screenshot of the overview of the connected learning modules.

Both implementations follow the concept of Semantic Specification Documents (SSD) that are self-contained and tool-agnostic [8], which are important properties to ensure an easy adoption in heterogeneous environments with a high variation of tools. PDF and web-based technologies are the document representation used, which are widely adopted and ELM is the backbone for the semantic representation, carrying the meaningful relations among the information elements.

V. Conclusion & Outlook

This contribution analyzed the current situation of the digital transformation within HEI with respect to the digital documentation and certification of learning outcomes (achievements). It demonstrated that a number of European standardization activities are already paving the way towards machine readable semantically enhanced specification documents with the European Learning Model (ELM) developed by the European Commission and with the Educational Verifiable Certificates (EVC) by the W3C. Similar activities can be found in other areas of the world, e.g., the Credential Transparency Description Language (CTDL) [21] and the OpenBadge Specification [22] in the US.

In Austria and many other European countries, the documentation of learning outcomes is already state of the art by supplying a Transcript of Records (ToR) and a Diploma Supplement (DS), which are often generated in an automated fashion, but still in the format of unstructured text. However, the aspect of explicit semantics and precise machine readability is not yet being used widely.

The case analysis and the prototype implementation demonstrated that the use of the standards is technically feasible without harming existing generation processes and that a number of stakeholders could benefit from self-contained, machine-understandable specification documents.

The next steps in this research efforts are a more extensive implementation for a number of course programs at the HEI of the author and the evaluation of the benefits for stakeholders for important use cases, such as the Recognition of Prior Learning (RPL) with different stakeholders. This will be the basis of an ongoing evaluation and improvement process that focusses on the ease of use and a beneficial user experience to ensure a good adoption rate to accelerate the digital transformation of HEI.

Acknowledgment

The author would like to thank the Tyrolean Science Fund (“Tiroler Wissenschaftsförderung”), which supported this research under grant number F.33280/6-2021 and the detailed and helpful comments of the reviewers of this contribution.

References

[1] European Commission, “The European Qualifications Framework (EQF) | Europass.” https://europa.eu/europass/en/europass-tools/european-qualifications-framework (accessed Aug. 01, 2023).

[2] S. Bohlinger, “Comparing Recognition of Prior Learning (RPL) across Countries,” in *Competence-based Vocational and Professional Education: Bridging the Worlds of Work and Education*, M. Mulder, Ed., in Technical and Vocational Education and Training: Issues, Concerns and Prospects. Cham: Springer International Publishing, 2017, pp. 589–606. doi: 10.1007/978-3-319-41713-4_27.

[3] European Commission, “Upcoming launch of the European Learning Model v3 | Europass.” https://europa.eu/europass/en/news/upcoming-launch-european-learning-model-v3 (accessed Jun. 15, 2023).

[4] “Verifiable Credentials Data Model 1.0.” https://www.w3.org/TR/vc-data-model/ (accessed Apr. 19, 2021).

 ISBN: 978-1-68558-108-4

[5] "Modeling Educational Verifiable Credentials." https://w3c-ccg.github.io/vc-ed-models/ (accessed Jun. 15, 2023).
[6] European Commission, "ECTS and Diploma Supplement Labels," *EACEA - European Commission*, Mar. 13, 2018. https://eacea.ec.europa.eu/sites/2007-2013/lifelong-learning-programme/ects-and-diploma-supplement-labels_en (accessed Apr. 09, 2021).
[7] European Commission, "Diploma Supplement | European Education Area." https://education.ec.europa.eu/education-levels/higher-education/inclusive-and-connected-higher-education/diploma-supplement (accessed Aug. 01, 2023).
[8] K. Böhm, "Towards Semantically Enriched Curricula as pre-built Information Spaces in Higher Education Institutions," in *ECKM 2021 22nd European Conference on Knowledge Management*, Academic Conferences limited, 2021, p. 71.
[9] "Seven Steps to a Solid Job Task Analysis," *EDSI*. https://www.edsi.com/blog/seven-steps-to-a-solid-job-task-analysis (accessed Aug. 01, 2023).
[10] BMBWF, "Electronic Signature." https://bmf.gv.at/en/topics/digitalisation/Digitised-Austria/Electronic-Signature.html (accessed Jun. 18, 2023).
[11] European Commission, "Diploma Supplement for stakeholders | Europass." https://europa.eu/europass/en/diploma-supplement-stakeholders (accessed Apr. 09, 2021).
[12] BMBWF, "Diploma Supplement." https://www.bmbwf.gv.at/en/Topics/Higher-education---universities/Recognition-of-qualifications/Diploma-Supplement.html (accessed Aug. 01, 2023).
[13] K. Böhm, "CloudEARTHi workshop on Agility and Semantic Structures for Scaffolding Academic Education, Part I and Part II." https://cloudearthi.com/asss-academic-education/ (accessed Sep. 05, 2023).
[14] L. Recalde, R. Navarrete, and L. Correa, *A Framework for data mining of structured semantic markup extracted from educational resources on University websites*. 2022. doi: 10.54941/ahfe1001745.
[15] Google, "Intro to How Structured Data Markup Works | Google Search Central | Documentation," *Google for Developers*. https://developers.google.com/search/docs/appearance/structured-data/intro-structured-data (accessed Aug. 01, 2023).
[16] R. Montti, "Google On Which Structured Data it Prefers: JSON-LD or Microdata?," *Search Engine Journal*, Mar. 11, 2019. https://www.searchenginejournal.com/google-structured-data-preference/297479/ (accessed Aug. 01, 2023).
[17] P. Wyatt, "PDF 2.0 Application Note 002: Associated Files," Oct. 26, 2018. https://pdfa.org/resource/pdf-2-0-application-note-002-associated-files/ (accessed Jun. 22, 2023).
[18] "PDF-LIB · Create and modify PDF documents in any JavaScript environment." https://Hopding.github.io/ (accessed Jun. 22, 2023).
[19] W3C, "JSON-LD 1.1," 2020. https://www.w3.org/TR/json-ld11/#embedding-json-ld-in-html-documents (accessed Apr. 13, 2021).
[20] K. Böhm, *Agility and Semantic Structures to Scaffold Modern Academic Education Supporting the Digital Transformation in HEI*. 2022.
[21] "Credential Engine Registry | Credential Transparency Description Language Schema Metadata." https://credreg.net/ctdl/terms (accessed Jun. 28, 2023).
[22] "Open Badges Specification 3.0 Candidate Final Public." https://1edtech.github.io/openbadges-specification/ob_v3p0.html (accessed Jun. 19, 2023).

 ISBN: 978-1-68558-108-4