CARLA 0.9: Now with more Analytics!

Carlos Natanael Lecona-Valdespino[1], Pablo Isaac Macias-Huerta[2], Rafael Cabañas-Rocha[2], Guillermo Santamaría-Bonfil[3]

[1] Universidad Panamericana, Mexico

[2] Instituto Politécnico Nacional, Escuela Superior de Cómputo, Mexico

[3] CONACYT-INEEL, Mexico

0245614@up.edu.mx, {pabloisaacmacias,cabanas.rocha.rafael}@gmail.com, guillermo.santamaria@ineel.mx

Abstract. E-Learning offers new mechanisms for easing interactions with learning environments. In an earlier work the Conversational Agent based in viRtual reaLity with Analytics (CARLA) was introduced. This is a chatbot that provides interaction within a virtual reality E-Learning platform through spoken dialogue. However, its knowledge base was restricted to a local database. To expand it, a search engine is integrated into the chatbot. Furthermore, to recommend new knowledge a recommender system can be employed which is dependable on a semantic similarity measure. Therefore, in this work we improve the CARLA framework by including a search engine for the dialogue manager, and provide support for recommendations. In particular, the recommender system is built upon the Normalized Web Distance which allows semantic comparison of words and concepts using a large corpus. Results using the Google and Wikipedia search engines show a high correlation against expert and common knowledge respondents, emphasizing that the Normalized Web Distance can be employed to suggest users new knowledge through the chatbot based on topic similarity and user queries.

Keywords: conversational agent, virtual reality, natural language processing, search engine, recommender system.

1 Introduction

E-learning technologies are of increasing importance in education and Virtual Reality (VR) environments are key in this area [5]. VR provides contextualized knowledge, improves learning through exploration and repetition, and has several degrees of freedom [7]. Furthermore, as Natural Language Processing (NLP) technologies mature, more conversational agents are being adopted within VR E-learning platforms. These have a wide range of applications including (but not restricted to) tutoring, question answering, language conversation practice, learning companions, and so on [4]. A new VR chatbot is CARLA: a Conversational Agent in viRtual reaLity with Analytics. It is based on Google® speech to text, Stanford CoreNLP, and Microsoft® Azure's text to speech technologies [5]. It is designed to eventually serve as a tutoring system situated in a virtual reality environment. In its current state, CARLA provides knowledge and support to a given user through an interchange of spoken utterances for manipulating objects and navigating the virtual world.

Nevertheless, CARLA was restricted to its local knowledge defined by experts in an internal database. Its maintenance and management is expensive, time costly, and error prone. In this sense, the Dialogue Manager (DM) of any chatbot that is constrained to such setup cannot provide unknown information nor discover new knowledge. To ameliorate this, a Search Engine (SE) is added to the DM of the chatbot [2]. This is a tool for improving the existing knowledge domain by adding search capabilities using a large text corpus such as the world wide web (www). Even though today SEs are optimized to show high relevance results, this optimization is oriented to the advertising of commercial results. Such bias will not be desirable for educational purposes since SEs may provide results with little to no pedagogic value [2].

To improve this, a Recommender System (RS) can be employed, with Collaborative Filtering (CF) being the most prevalent type [3]. It clusters users or items by comparing user-user or item-item similarity, respectively. However, CF based only in local resources has two main drawbacks: it requires a large amount of data to avoid the sparsity in the similarity matrix, and it suffers from cold-start (new items which lack of knowledge). To avoid this, the Normalized Web Distance (NWD) has been employed [3]. The NWD can be used to identify similarity between topics, either in the local resources or collected from user queries, done through the SE which has a readily available large text corpus such as the www [6].

Therefore, in this work the CARLA framework is improved by 1) integrating a SE such that unknown knowledge in the current database can be found and provided to users on request, and 2) building a recommender system to suggest new learning resources based on items' semantic similarity given by the NWD. In particular, we compare the chatbot recommendations' ability against human expert and common knowledge respondents. For this, 20 triplet concepts (10 specific and 10 common topics) are compared using an online survey. In parallel, Google and Wikipedia SEs are used to calculate the NWD over the 20 triplets.

The results show that there is a high correlation between the scores obtained by humans and the NWD over the 20 triplets, specifically the ones obtained by the Wikipedia search engine. The rest of the paper is organized as follows: section 2 presents the materials and methods; section 3 presents the experimental setup and the results of the experimentation; section 4 presents conclusions and discusses future work.

2 Materials and Methods

CARLA is an embedded chatbot framework for Virtual Reality (VR) using low-cost or free technologies. As any other chatbot, its main components are the Natural Language Understanding (NLU), the DM, and the Natural Language Generator (NLG).

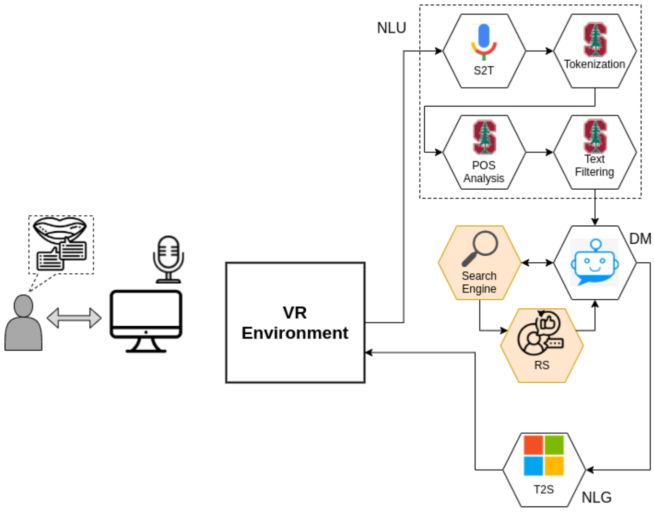

Fig. 1. CARLA framework with the DM and SE.

To improve the DM a search engine is added. This SE is used to create a recommender system based on the NWD. In Fig.1, we present the structure of the improved CARLA framework where the highlighted components represent the new additions.

2.1 Search Engine

A DM and SE are the two main components of any chatbot [2]. However, since CARLA was limited to its hard-coded database, it was necessary to implement a SE that could provide the user information unknown to CARLA. Also, the chatbot needs to be able to find answers in structured data and manipulate them in order to present them to the user. In this manner, when asking CARLA something, the DM checks if it has that information internally and if not, it asks the user if they would like to search for the information on a SE such as Google or Wikipedia; if the user confirms, CARLA will carry out the request and extract the first result given by the SE.

If the results obtained from the SE do not fully match what the user requested, they will be asked if they would like to hear the extract from the page anyway. Technically, the DM gets the information by requests to an API web service such as Google Search or the MediaWiki Action. Both are RESTful web services based on the *OpenSearch* bundle, a technologies collection which allows publishing search results in other websites and gathering results from multiple sources [1].

For example, the MediaWiki Action API allows searching and fetching information from any domain created for a Wikimedia project, such as the categories to which it belongs, images it contains, descriptions, page extracts, and even allowing page operations like creating or deleting a page. The Snippet 1.1 shows the results obtained by the API of a full-text search of *Computadora* from Wikipedia in Spanish with a

GET request to[4] which contains information such as the number of pages that contain the term, as well as information from all of them, including the title, its word count or an excerpt from every page.

Snippet 1.1. API results coded into json format


```
{
    "batchcomplete": "",
    "continue": {"sroffset": 10,"continue": "-||"},
    "query": { "searchinfo": {"totalhits": 13841},
        "search": [{
                "ns": 0,"title": "Computadora",
                "pageid": 8985,"size": 33529,
                "wordcount": 4085,
                "snippet": "La <span class=\"searchmatch\">
    computadora</span>[1]\u200b[2]\u200b (del ingl\u00e9s:
    computer y este del lat\u00edn: computare,[3]\u200b
    \u2018calcular\u2019), tambi\u00e9n denominada <span
    class=\"searchmatch\"> computador</span>[4]\u200b[1]
    \u200b u ordenador[5]\u200b (del",
                "timestamp": "2021-08-17T22:41:40Z"},
            {
                "ns": 0, "title": "Computadora port\u00e1til",
                "pageid": 9086,"size": 23757,
                "wordcount": 3222,
                "snippet": "Se denomina <span
    class=\"searchmatch\"> computadora</span> port\u00e1til,
    <span class=\"searchmatch\">computador</span>
    port\u00e1til u ordenador port\u00e1til, a un
    determinado dispositivo inform\u00e1tico que se puede
    mover o transportar con",
                "timestamp": "2021-07-30T20:29:57Z"
            }]}
}
```


2.2 Normalized Web Distance for Powering a Recommender System

A universal similarity measure is the Normalized Information Distance (NID), which allows identification of the similarity between two objects based on Kolmogorov complexity. However, the Kolmogorov complexity is not directly computable, thus, its practical application is cumbersome [8]. A measure which is based on the NID is the so called NWD. This measure employs the www and SE to approximate the semantic similarity (circumventing the need of estimating the Kolmogorov complexity) between two queries or concepts.

In particular, the NWD excels in determining the semantic similarity for *non-literal objects* that do not contain information within themselves [6]. Take the concept *wind* for example, the sequence of letters or the word itself does not contain any information

[4] https://es.wikipedia.org/w/api.php?action=\\query&list=search&srsearch=Computadora

regarding what *wind* stands for. Formally, if x and y are two different concepts within the set W of all web pages in the www corpus, i.e., $x, y \in W$, then, the NWD between these two is given by

$$NWD = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}, \tag{1}$$

where $f(x)$ and $f(y)$ are the number of pages containing term x and y respectively, $f(x,y)$ is the number of pages containing both x and y, and N is a large constant corresponding to the total number of indexed pages. The NWD is roughly between 0 and infinity, obtaining values close to zero for objects which are very similar, and one or larger for very dissimilar objects. In some cases the NWD can obtain values below zero if $f(x,y) > \max\{f(x), f(y)\}$, or infinite if objects do not appear together on the same page, but do occur separately.

Also, the NWD is considered a relative similarity measure between two concepts (due to its dependence to the SE) rather than a similarity metric [6]. On the other hand, for a proper usage of the NWD for powering a recommender system, the proper SE needs to be selected. It is the case that many SEs are optimized for advertisements. If we employ a biased SE to calculate the NWD, we can hinder new knowledge suggestions to a given user. Therefore, to evaluate this hypothesis we employ Google and Wikipedia as SEs. The first is the most popular SE, whereas the second is a free, multilingual, online encyclopedia written mainly by volunteers.

3 Experimentation

In the following, we analyze the level of knowledge agreement between humans and the NWD using two different SEs. First, we present the details of the survey answered by human participants. Then, we present the analysis of the results obtained by the two SEs, and the comparison with the survey results.

3.1 Survey and NWD Calculation

We elaborated a survey[5] with triplets of concepts where participants are requested to determine which of two options is the more closely related to a given concept. In the first part of the survey users informed consent and personal data is gathered. The latter is conformed by gender, age group, educational level, and country. It is worth mentioning that for the educational level we consider the following groups: primary, secondary, high school, undergraduate, and postgraduate.

The survey was responded 130 times by 56% male, 43% female, and 1% did not specify. Regarding the age group, 50% of the participants were older than 30, 44% were between 18 and 29 years old, and only 6% were between 12 and 17 years old.

Regarding the educational level, 38% of the participants were high school students, 36% of the participants had some postgraduate degree, 25% were undergraduate students, and only 1% were secondary school students. Regarding the nationality of

[5] https://forms.gle/sDBZdfVFgRTGiPXc7

the participants, 98% were Mexicans, and 2% from other countries. The second part of the survey assesses 20 triplets' (e.g., (A,B,C)) kinship using a Likert scale. This scale presents a given concept (e.g., A), and asks human participants to assess its similarity with two others (e.g., B and C), with one of them, for a physical or chemical reason, being more related to the main concept. Some triplets show highly related concepts, while others have concepts with high similarity to one another.

The intention is to determine the level of *agreement* between users and the NWD. The scale ranges from 0 to 10 (from left to right), and the concepts are presented in opposite sides. In this sense, if the human chooses 0, concept B is the most related to A; if instead human picks 10, concept C is the most related to A. If both concepts B and C are equally similar/dissimilar to A, then, humans need to pick the middle value of the scale (i.e., 5). The triplets are composed by 10 concepts in renewable energy and 10 from general domain. For the former, we only consider the answers of participants with some postgraduate degree given that more advanced knowledge is required; an example of this type of triplet is (*energía fotovoltaica, aerogenerador, panel solar*) (translated to English: photovoltaic energy, wind turbine, solar panel).

For the latter, the responses of all participants is considered; an example of this type of similarity assessment is $(manzana, pera, libro)$ (translated to English: apple, pear, book). To determine the similarity assigned by human participants, for each item of the survey the mean value of the responses is calculated. On the other hand, given a triplet of concepts we use the NWD and two different SEs, namely Google (NWD_G) and Wikipedia (NWD_W), to determine the similarity between the main concept and the other two. For both cases Eq. 1 is employed to calculate the similarity between the concepts i.e., $NWD_{G,W}(A,B)$ and $NWD_{G,W}(A,C)$. For each SE two results are obtained, thus, to determine the most related concept the lowest value (i.e., the most semantically related) is selected.

3.2 Similarity Comparison and its Use for Recommendation

Results about renewable energy and the general domain concepts for the survey and the $NWD_{G,W}$ are presented in Table 1 and Table 2, respectively. The most semantically related pair of concepts are highlighted in bold. First, observe that in the case of the survey with human participants, and for most of the items, a high polarization is shown i.e., tending to just one or the another concept. In contrast, results for the $NWD_{G,W}$ show less polarization. This is very interesting, since humans tend to provide more polarized answers, presumably because their knowledge is highly contextualized.

For example, for the last triplet of Table 1 which translates to (coal, mine, dam), humans with specialized knowledge provide a value of 0.15 which depicts a high similarity between *coal* and *mine*. In contrast, the NWD_G achieved a 0.69 and 0.55 for the NWD_W, which highlights similarity but not quite as sharply as human participants considered. Nevertheless, *for a given concept and two options, do humans and the NWD agree in picking the most similar concept?*

The answer is yes, with a linear correlation of $\rho = 0.99$ for the renewable energy concepts and $\rho = 0.95$ for the general domain concepts, regardless of the SE (with a $\rho = 1$ between the NWD_G and NWD_W). These results are encouraging since the NWD can be used to suggest new, semantically similar concepts or topics to a user given their history of already-reviewed topics. On the other side, regarding the agreement between survey participants' mean scores and the raw score obtained by the heuristic of picking the lowest NWD value, for the renewable energy domain, the NWD_G obtained a higher correlation $\rho(NWD_G) = 0.93$ in comparison to $\rho(NWD_W) = 0.9$, with a level of agreement between the two of $\rho(NWD_{G,W}) = 0.89$.

For the general domain concepts, the agreement with humans is $\rho(NWD_W) = 0.94$ and $\rho(NWD_G) = 0.89$, and an agreement between the two SEs of $\rho(NWD_{G,W}) = 0.87$.

These results are counterintuitive given that one would expect that the NWD_W, based on an encyclopedia, should be biased towards domain specific knowledge, whereas the NWD_G should work better for general domain concepts given its bias to the advertising optimization. This might be caused by the Wikipedia SE being constrained to its Spanish version whereas Google SE is open to any language.

Table 1. Survey and NWD results for renewable energy concepts.

Triplet	Survey	$NWD_G(A,B)$	$NWD_G(A,C)$	$NWD_W(A,B)$	$NWD_W(A,C)$
(energía fotovoltaica, aerogenerador, panel solar)	9.59	0.7	**0.01**	0.33	**0.22**
(energía geotérmica, volcán, marea)	0.34	**0.6**	0.77	**0.16**	0.76
(energía eólica, aerogenerador, panel solar)	0.13	**0.31**	0.47	**0.13**	0.71
(almacenamiento de energía, baterías, cable)	0.13	**0.55**	0.86	**0.44**	0.75
(combustible fósil, renovable, no renovable)	9	0.44	**0.33**	0.43	**0.22**
(electricidad, televisión, estufa)	0.88	**0.54**	0.63	**0.56**	0.69
(térmico, calor, luz)	0.54	**0.42**	0.66	**0.49**	0.91
(energía no renovable, energía hidráulica, central nuclear)	7.56	0.27	**0.05**	0.71	**0.32**
(energía renovable, gasolina, granja eólica)	9.88	0.77	**0.11**	0.58	**0.43**
(carbón, mina, presa)	0.15	**0.69**	0.75	**0.55**	0.92

Table 2. Survey and NWD results for general domain concepts.

Triplet	Survey	$NWD_G(A,B)$	$NWD_G(A,C)$	$NWD_W(A,B)$	$NWD_W(A,C)$
(manzana, pera, libro)	0.54	**0.42**	0.68	**0.55**	1.19
(nube, agua, algodón)	1.83	**0.36**	0.72	**0.74**	0.76
(experimento, laboratorio, datos)	2.85	**0.45**	0.53	**0.62**	1.15
(plato, cuchara, pelota)	0.51	**0.45**	0.82	**0.90**	1.12
(pepino, sándwich, ensalada)	8.91	0.6	**0.34**	0.67	**0.37**
(robot, ciencia, tarot)	0.35	**0.27**	1.57	**0.75**	0.8
(vehículo, movimiento, mar)	0.84	**0.56**	1.02	**0.84**	1
(sillón, silla, mesa)	1	**0.5**	0.7	**0.6**	0.73
(zapato, pie, mano)	0.29	**0.63**	0.76	**0.82**	1.1
(lentes, vista, oidos)	0.32	**0.71**	0.74	**0.72**	1.07

4 Discussion and Conclusion

Conversational Agent in viRtual reaLity with Analytics (CARLA) is a spoken chatbot built upon natural language tools such as Google NLU, Stanford CoreNLP library, and Microsoft NLG. However, CARLA was restricted to its local, limited, expert-curated knowledge database, which limited its pedagogical usage and made its maintenance expensive.

To ameliorate such burdens, we improve CARLA by adding to its dialogue manager a SE. This allows the exploration of a large text corpus such as the www, for knowledge not already available to CARLA. Moreover, the usage of a SE allows the DM to estimate the NWD which can be used to assess semantic similarity between topics.

This feature can be used by a recommender system to suggest new topics to a user, related to the knowledge already explored by them. Therefore, to assess the NWD potential for powering a recommender system, we compare the similarity assigned to triplets of concepts between humans and the NWD obtained using two different SEs (Google and Wikipedia). The agreement between humans and the NWD for the survey items as measured by the linear correlation is, in general, $\rho \geq 0.9$.

These results are encouraging since the NWD can be used to suggest new, semantically similar concepts or topics to a user given their history of already reviewed topics. Such topics can be located within CARLA's local knowledge base or in the pervasive www. However, which SE must be used to suggest new semantic similar topics is still an open question.

Therefore, future work will involve the construction of a semantic concepts network, built upon the NWD, which will be used by the chatbot to walkthrough a given knowledge area. Then, the users will be requested to evaluate the concepts roadmap using different SEs.

References

1. Clinton, D.: Spezifikation opensearch 1.1 draft 3 (2007), http://www.opensearch.org/Specifications/OpenSearch/1.1
2. Galitsky, B.: Developing Enterprise Chatbots. Springer International Publishing (2019)
3. Huang, T.C.K., Chen, Y.L., Chen, M.C.: A novel recommendation model with Google similarity. Decision Support Systems 89, 17–27 (2016)
4. Kerry, A., Ellis, R., Bull, S.: Conversational agents in e-learning. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence. pp. 169–182. Springer (2008)
5. Macias-Huerta, P.I., Santamarıa-Bonfil, G., Ibanñez, M.: Carla: Conversational agent in virtual reality with analytics. Research in Computing Science 149(12), 15–23 (2020)
6. Santamaría-Bonfil, G., Díaz-Rodríguez, H.D., Arroyo-Figueroa, G., Batres, R.: Knowledge modelling for ill-defined domains using learning analytics: Lineworkers case. In: International Conference on EUropean Transnational Education. pp. 409–418. Springer (2020)
7. Santamaría-Bonfil, G., Ibáñez, M.B., Pérez-Ramírez, M., Arroyo-Figueroa, G., Martínez-Álvarez, F.: Learning analytics for student modeling in virtual reality training systems: Lineworkers case. Computers & Education p. 103871 (2020)
8. Vitanyi, P.M.B., Balbach, F.J., Cilibrasi, R.L., Li, M.: Normalized Information Distance, pp. 45–82. Springer US, Boston, MA (2009)