Construction of a Knowledge Map-based System for Personalized Second Language Learning

LOUIS LECAILLIEZ[†1] BRENDAN FLANAGAN[2]
HIROAKI OGATA[2]

Abstract: A great number of Computer-Assisted Language Learning (CALL) systems have been developed, focusing on one particular language skill and a limited range of user expertise. In this paper, we propose the use of a knowledge graph that contains both lexicon and grammatical data that serve as the basis of learner's model. By being populated with event log data from external sources, the graph could serve as a user model for client systems in a lifelong learning perspective. The knowledge map is exposed as a service that can be interfaced with existing CALL and learning analytics tools. Beyond the direct display of the knowledge map to the learner that most existing systems provide, we investigate text-based use cases, such as: recommending grammatical concepts with examples tailored to the user's current knowledge, and quiz generation that feeds the system's feedback loop.

Keywords: Concept Map, Knowledge Graph, User Model, Second Language Education, ICALL

1. Introduction

Since the advent of personal computer, a great number of Intelligent Computer-Assisted Language Learning (ICALL) and Intelligent Tutoring Systems (ITS) have been developed [1]. These systems are typically targeted towards teaching one or two language skill among reading, writing, listening and speaking [2]. ICALL perform their duties via techniques such as natural language processing (NLP), user modeling, or natural language generation [2].

In particular, systems that use user modeling try to keep enough information about the learner knowledge to adapt their difficulty level. There are however three issues that can be drawn from the literature. First, Slavuj [3] says that ITS are "over-restricting the learning domain" and "focus on a single linguistic skill". Secondly, each system maintains its own domain and user modeling that is not exportable to another system. Finally, [3] notes "an inability of an ITS to cater for learners with different levels of language proficiency."

This first and third issues imply that no single existing CALL cater to all the linguistic needs of a student, moreover so during its progression from beginner to advanced. The second issue is a problem as well in the perspective of a lifelong language education: each time a learner changes institution or platform, its user model is lost for the ICALL newly used. This is why we raise the question: how can the learner knowledge of a second language be modeled in a sufficiently fine-grained manner to support multiple skills for multiple CALL systems? We believe graphs is an answer and will investigate their use as user model.

The rest of this paper is organized as follows: Section 2 will review related work, i.e. occurrence of human knowledge structured as graph. Section 3 will discuss a graph model for taking into account both lexicon and grammar points of a language. Section 4 will describe the architecture of a computing system based on the graph discussed previously and will give use case examples. Sections 5 and 6 respectively discuss the presented system and conclude the paper.

2. Related Work

Related work, i.e. the modeling of human knowledge as a graph is dispersed under multiple names in different fields. We identified three of them: *concept map* in education, *knowledge graph* in semantic web and *lexical network* in lexicography. These three concepts have different goals and applications, but they all share a graph-based structure.

2.1 Concept Maps

Concept maps are a pedagogical tool born from the need of representing hierarchical conceptualization of scientific concepts [4]. A concept map (Cmap) is formed by words denoting concepts that are linked together by a directed, labelled relationship. It can be used in a variety of ways: usages in education include student's reflection on his knowledge, testing a learner, or comparing the content of two curricula [4].

Not surprisingly, Cmaps have also been employed for various aspects of second language (L2) teaching. For instance, Rhoii and Sharififar [5] compared Cmaps to rote learning for vocabulary acquisition. Dias [6] researched the improvement that Cmaps can provide to L2 English reading. Lee [7] and Ojima [8] investigated the use of Cmaps in respectively Korean and Japanese writing. Cmaps are versatile and can be used for multiple language skill. Note that the cited studies do not make use the use of computers.

2.2 Knowledge Graph

The term knowledge graph was coined and popularized by Google in 2012 to describe their internal use of semantic knowledge [9]. The word has then been appropriated by the Semantic Web community to refer to knowledge databases implemented with RDF (Resource Description Framework). However, there is no to these days any formal definition or agreement on the exact meaning of the word.

1 Graduate School of Informatics, Kyoto University, 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
2 Academic Center for Computing and Media Studies, Kyoto University, Yoshida-Nihonmatsu, Sakyo-ku, Kyoto 606-8501, Japan.
† louis.lecailliez@outlook.fr

While Cmaps are typically used by a human to get a quick overview of a single subject, knowledge graphs are built to encompass very broad human knowledge, possibly dispersed on various fields and domains, in a way that can be challenging even for a computerized system [10].

2.3 Graph Lexical Resources

The most popular contemporary lexical resource structured as a graph is Wordnet [11], a large lexical database of English words. Wordnet (WN) is made of nouns, verbs, adjectives and adverbs linked in a set of cognitive synonyms called a synset. Each synset expresses a distinct concept. Synsets are linked together mainly on the basis of semantic relationships such as synonymy or hypernymy. Wordnet structure is based on psycholinguistic theories about the organization of lexical information in the human brain [12].

Multiple wordnet projects have spanned in other languages: French [13], Japanese [14], Czech [15] are examples of this development. Internationalization attempts can be divided in two approaches: merging or extending [13]. The *extend* pattern consist of automatically translating the English Wordnet and then humanly correcting the whole or a part of the resulting network. It is a common approach because of the size of the original WN.

Wordnet could be an interesting base for a system trying to keep track of the user knowledge of a language. There are however the following issues. Firstly, the network only contains lexical information, and no other concerns of language learning such as grammar. Secondly, in international versions of WN the translation process generated errors which is a serious problem for a teaching tool. Finally, while the structure of WN built around hypernymy and hyponymy can be put in correlation with the mastering of a language [16], this structure may not be the most useful at the beginner level, thus contradicting the stated goal of a lifelong educative solution.

2.4 Proposed Graph

In this work, we use different aspects of the three graph-based objects reviewed previously. Specifically, we will create a graph that is akin to a concept map in that it includes concepts such as grammar points to be learned. But in contrast to Cmap that are designed to be handled by human and thus have a fairly limited size, our graph will contain a huge number of nodes representing lemmas and grammar points of the target language. The focus being on language learning, general concepts found in knowledge graph will not be present, except for concepts that have a second language acquisition significance such as cultural aspects concerning the people using the language natively.

Another key difference with existing works, that aim to objectively describe knowledge (be it general, domain specific or lexical) is the inclusion of user data in the graph by the mean of dedicated nodes. The resulting graph is thus not purely a concept map or a lexical network but a heterogeneous graph that describe the hypothetical knowledge state of multiple users.

3. Graph Modeling

3.1 Context

As discussed earlier, wordnets for languages other than English mostly come from translated content and thus are unsuitable for direct use as the basis of the user model. However, a graph structure is interesting for multiple reasons: (1) it may be close to the actual psycholinguistics representation of knowledge and therefor is adapted to actual knowledge representation, (2) it is trivially transformable to a concept map (for which there is proven usefulness in education) for display and interaction by a student or teaching staff, (3) it is a well-supported both from the theoretical point of view and from the data storage tooling availability and (4) it is easily extensible with external data.

There exists a variety of algorithms to process graphs: the Wikipedia page on the topic lists 113 of them[a]. Some of the most well-known ones are the A* search algorithm to find paths between two nodes, the Kruskal algorithm to create a minimum spanning tree, and Dijkstra algorithm for shortest path finding. So, if an educational task can be expressed as the result of a graph algorithm then there is an already well-known way to implement it. For instance, the “Louvain algorithm” [17] could be used to detect groups of users of a similar level in a very large graph.

3.2 Knowledge Map Structure

For the proposed system to be adaptable to multiple tasks and languages, it must be usable with graphs of different structure. The current core graph schema is described below, but as the system is still being in development, this may not reflect exactly the final structure.

The system is architectured around a graph that acts as the content to be taught or learned and as a repository of the knowledge of every user. This graph is created with the open-world assumption: something not in the graph may exists and could be added to it later on. Users are added to the graph as a nodes and links are formed to words and concepts as they learn new material. Edge annotations are used to store additional information such as the source that triggered the creation of link or the confidence level associated to it. More details about the whole architecture is given in section 4.

The core graph is made of three mandatory node types. The main one is the high-level grammatical concept to be taught e.g. expression of the future, conditional or plural. A concept can be subdivided in more granular sub-concepts; in a Cmap we developed for Sumerian, genitive constructs are split in 4 kinds of construction with different structure and usage. Another mandatory type of node is the example and words node. A word node adds the ability to keep track of the learner’s vocabulary knowledge. This is useful for making smart recommendations. The example nodes are required to create connectivity between words in the graph because in contrary to Wordnet there is no predefined relationship between them.

Word, Example and Concept nodes form the core of the language description to which user data can be associated. Additional nodes can be used to store more information about a

a https://en.wikipedia.org/wiki/Category:Graph_algorithms

curriculum content or retaining information from the source that served to create the Cmap. Figure 1 presents the schema of a Cmap used for Standard Chinese teaching. It was built from pages of the Chinese Grammar Wiki[b] (blue nodes) and vocabulary lists compiled for the Hanyu Shuiping Kaoshi standardized test (nodes in red). Additional node types encode notably relationship with page title, Common European Framework of Reference for Languages (CEFR) level and Chinese characters. For simplicity of display, edge identifiers have been masked.

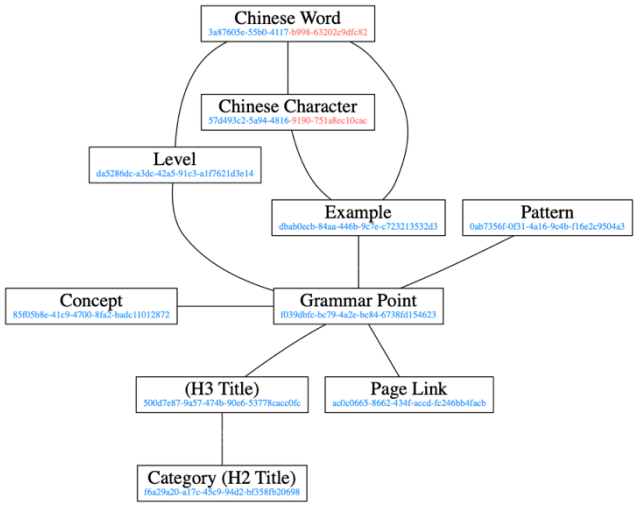

Figure 1 Schema of a graph for Chinese learning

3.3 Extensibility and Interoperability

One of the key advantages of using a graph structure for data modeling is the ease with which the database can be extended without requiring migration. In particular adding a new data type is easy and doesn't require change to any existing data except for the creation of the link to existing node. In comparison, adding new data to a relational Database Management System (DBMS) require adding additional columns and tables to link related existing to new data.

In our case, we have to deal with language: it features hundreds of thousands of lemmas. Even with a set of controlled size (e.g. the core 3000 words of a language), adding enough additional data for some scenario can be challenging. For example, let's consider providing a definition to words present in the graph. This could be done easily with a graph: the definitions are added to the graph in their own new nodes and links are created to the existing word. In particular, the use of a graph representation of a language lexicon and grammar rules mean it could be extended by linking nodes to existing resources from the Linguistics Linked Open Data (LLOD) cloud [18].

Knowledge maps developed for the system could be exposed as linked data or shared as a file for other researchers to use. Combination, extension and remixing of models is interesting for lowering the cost and efforts of new projects or to test hypothesis in new settings.

4. Architecture

4.1 Overview

A significant part of CALL systems is web-based portals that provides an educational service to learners. They are implemented with a web server serving dynamically generated webpages while transferring data back and forth a relational DBMS. This straight-to-the-point approach mean that the user model, the learning data and all related data are stored in a siloed database with a schema fitting the need of only this application.

In recent years, Learning Record Store (LRS) have emerged from the need to consolidate and analyze learning interactions from multiple applications. LRS can provide better insights thanks to their ability to aggregate data from multiple services. The decoupling of the student logged actions and the CALL system is an interesting idea. We want to push it further and also decouple the user model from the CALL system.

The platform we propose is independent of any given existing tool. It is designed to acquire data from existing systems or from a LRS. Data is inputted into the system through a JSON REST API. When a LRS is present in the infrastructure, data is instead pulled periodically by the KMS. This exception arises because a LRS does not specify a way to relay events.

The figure 2 shows the high-level architecture of the project. On top, a list of CALL/Mobile-Assisted Language Learning (MALL) or other learning systems send data about user knowledge and progression to the Knowledge Map Server (KMS). Data is exchanged in JSON which is now the industry standard for web service communication. The KMS persists its data in an external graph database.

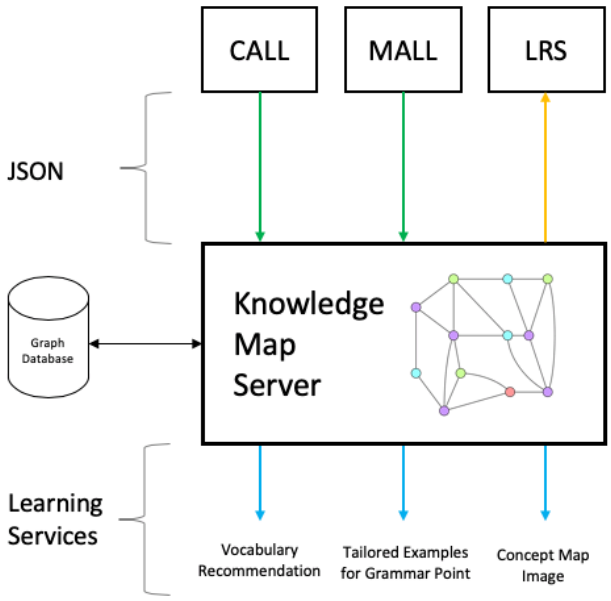

Figure 2 Architecture of the proposed system

The knowledge map server stores user model for multiple CALL systems. With this architectural choice, it is possible to interface or develop CALL specialized in a particular language

b https://resources.allsetlearning.com/chinese/grammar/Main_Page

skill while keeping the user model in a single unified model. However, to be useful the KMS must also be able to provide services based on the user model to clients CALL.

4.2 KMS Output

One obvious usage of the KMS mentioned in Figure 2 is export of the known vocabulary list of a learner. A client platform could then use this information to generate a quiz. More advanced queries could provide lists filter on a combination of criterions, such as the learning date of a word. A more interesting example case is the creation for a given level of a vocabulary list unknown to the user but mastered by the 5 learners that have the closest knowledge profile.

In response to the first issue raised in the introduction, the integration of both vocabulary and grammar knowledge in the same graph could open the way to scenario such as generating a grammar lesson for a grammar point where example sentences contain only known vocabulary. Likewise, similar outputs could be generated on this template by varying the number of unknow words in example sentences: none, exactly one, at least two, etc. In addition, the corresponding vocabulary could be generated alongside the lesson.

Finally, the last example we will give is the direct generation of an image of a knowledge map of the user knowledge, or a subgraph of it centered on a particular node.

4.3 Real World Use Case

The architecture presented before is abstract and generic. We will now present a real-world use case to show how the system could be used to support different aspects of learning. The KMS is integrated with three existing systems. The first one is BookRoll, a web-based e-book reader system that logs reading activity [19]. In our case, BookRoll is used only for data acquisition. The second integrated system is SCROLL [20], a learning platform that allows students to take notes and to play quiz. SCROLL is here both a data source and a client system of KMS. The third system present in the demonstration is a Dashboard for visualizing concept maps [21]. The dashboard display KMS output and doesn't transfer data to it. Figure 3 details data flow between these platforms.

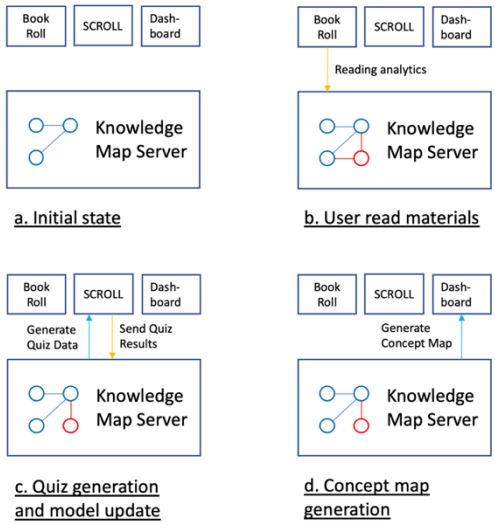

Figure 3 Integration of a KMS with existing platforms

At first (a), the model for the user is empty and the KMS contains a Cmap based on learning material and curriculum for the target language. After reading the first lesson (b), a user node is created in the KMS and nodes present in the lesson are considered known. When the user record new word via SCROLL, the corresponding nodes are marked as known. Through SCROLL, the user can generate a quiz (c) for the last 10 words added to its model. Once the quiz is done, the results are sent to the KMS. Confidence associated with the nodes present in the quiz are updated (incremented, decremented or set to zero) given the score results (c). Note that the sore update is dependent of an algorithm and no algorithm could fit every use case. As such, every data transferred to the KMS is written to write only log file that could later be used to construct a different user model from the same data. Finally, the user could view and navigate in a concept map via the dashboard portal (d).

5. Discussion

The presented system has the same flaw as every CALL/ITS system: it is depending on its inputs to evaluate correctly the user knowledge. Language acquisition made outside of the systems inputting data in the KMS is not tracked. However, the issue is alleviated in comparison to existing system by the fact that it aggregates the data from multiple CALL. It should thus be able to draw a clearer picture of the user level.

6. Conclusion

We presented a graph-based software system to track language knowledge of a student. It is aimed to solve three issues present in existing CALL/ITS systems. First, the versatility of its underlying graph model allows modelization of different language skills and user knowledge of vastly different level. Secondly, it consolidates data from multiple platforms which allow better student modelling as well a data portability and interoperability with ICALL systems. Finally, it provides educational service such as the generation of quizzes, concept maps or customized lessons that can be integrated in existing systems or used in brand new platforms.

Acknowledgments

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innovation Program on AI and Big Data 18102059-0.

Reference

[1] Chapelle, C. A., Sauro, S. (Eds.). (2017). *The handbook of technology and second language teaching and learning*. John Wiley & Sons.

[2] Gamper, J., Knapp, J. (2002). *A review of intelligent CALL systems. Computer Assisted Language Learning*, 15(4), 329-342.

[3] Slavuj, V., Kovačić, B., & Jugo, I. (2015, May). Intelligent tutoring systems for language learning. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 814-819). IEEE.

[4] Novak, J. D., Canas, A. J. (2007). Theoretical origins of concept maps, how to construct them, and uses in education. *Reflecting*

Education, 3(1), 29-42.

[5] Khoii, R., & Sharififar, S. (2013). Memorization versus semantic mapping in L2 vocabulary acquisition. *ELT journal*, *67*(2), 199-209.

[6] Dias, R. (2010). Concept map: a strategy for enhancing reading comprehension in English as L2. In *Concept Maps: Making Learning Meaningful*. Eds Sánchez, J., Cañas, A. J., & Novak, J. D., 29-33.

[7] Lee, Y. (2013). Collaborative concept mapping as a pre-writing strategy for L2 learning: A Korean application. International *Journal of Information and Education Technology*, 3(2), 254.

[8] Ojima, M. (2006). Concept mapping as pre-task planning: A case study of three Japanese ESL writers. *System*, 34(4), 566-585.

[9] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3), 489-508.

[10] Zeng, K., Yang, J., Wang, H., Shao, B., Wang, Z. A Distributed Graph Engine for Web Scale RDF Data. In PVLDB 2013: *Proceedings of the 39th international conference on Very Large Data Bases*, Riva del Garda, Trento, August 26-30, 265-276.

[11] Miller, G. A. (1995). Wordnet: A Lexical database for English. *Communications of the ACM* Vol. 38, No. 11, 39-41.

[12] Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

[13] Sagot, B., Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*.

[14] Pala, K., Smrž, P. (2004). Building czech wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2), 79-88.

[15] Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K. (2008). *Development of the Japanese WordNet*.

[16] Crossley, S., Salsbury, T., McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, 59(2), 307-334.

[17] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

[18] Chiarcos, C., Hellmann, S., & Nordhoff, S. (2011). Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. *TAL*, 52(3), 245-275.

[19] Flanagan, B., Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal*, *10*(4), 469-484.

[20] Ogata, H., Li, M., Hou, B., Uosaki, N., El-Bishouty, M. M., Yano, Y. (2011). SCROLL: Supporting to share and reuse ubiquitous learning log in the context of language learning. *Research & Practice in Technology Enhanced Learning*, 6(2).

[21] Flanagan, B., Majumdar, R., Akçapınar, G., Wang, J., Ogata, H. (2019) Knowledge Map Creation for Modeling Learning Behaviors in Digital Learning Environments. *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge (DC@LAK19)*.