ISSN 1870-4069

A Semantic Proposal for Semiautomatic Corpus Creation in the Pedagogic Domain

Yuridiana Aleman, María Somodevilla, Darnes Vilariño

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla, Mexico

{yuridiana.aleman,mariajsomodevilla,dvilarinoayala}@gmail.com

Abstract. In this research is introduced a methodology for the ontology automatic construction for the pedagogical domain based on the definition of two lexical resources: a domain corpus and, the use of a domain dictionary. In this research, the pedagogical corpus was building manually, where aspects such as learning strategies, intelligences types and learning styles were included. After that, the corpus is processed for extracting a list of concepts linked to the domain predefined classes. Each concept is searched in five dictionaries, including synonyms. This process is automatically performed, removing close words, words with low frequency. As result of this process, is building a domain dictionary which include concepts related with each domain class. The preliminary results show an affinity between the corpus and the dictionary, which is a very important resource for an ontology definition process. As future work, these resources will be used for detect classes in a ontology learning process for the pedagogical domain using machine learning techniques.

Keywords: Dictionary, synonyms, pedagogy domain, class detection, learning styles, intelligences types, learning strategies, corpus.

1 Introduction

In recent years, the available information has increased exponentially and the classic techniques of information retrieval cannot resolve problems such as semantic questions and language interpretation. Ontologies can be used for purposes such as structure knowledge in taxonomies, vocabulary manage, natural language processing applications, searches, recommendation systems, and e-learning among others [5]. The ontology learning process integrates the class detection, creation, populate and evaluation of ontologies. This process is applied in several domains focusing on one or more steps such has education, tourism, financial, among others ([9,6,7,16]).

Ontology is a formal, explicit specification of a shared conceptualization. Their classes, relationships, constraints and axioms define a common vocabulary to share knowledge [11]. The ontology learning process needs resources with specific characteristics and compatible with the domain to research. This paper is focused on the previous step of ontology learning: resources construction for class detection in the future. The resources are a dictionary with principal concepts and a corpus with papers published, the research will be realized for pedagogical domain in Spanish language. Pedagogical domain is extensive, thus the research is focused in the creation of support tools for the teachers in a class. Three topics was research: learning styles, intelligences types and learning strategies in the class.

The article is organized in 7 sections following described. Section 2 introduces the problem as well as the relevance to the selected domain. Section 3 presents the proposed methodology for the resources construction, and the sections 4 and 5 describes the resources. Preliminary results of quantitative analysis is shown in section 6. Finally, section 7 presents conclusions and future work for the research.

2 Related Work

In this section, the works about class detection and corpus created are analyzed.

Ontologies for the Use of digital learning Resources and semantic Annotations on Line (OURAL) project is presented in [9], the project includes people from several disciplines (educational science, computer science, and cognitive psychology) building e-learning services. The authors present the extracted class using Natural Language Process techniques in unstructured texts about learning situation. Educational domain was also analyzed in [6], but its application was into Chinese language. The authors analyzed three features of Chinese language: Coupling, Domain Relevancy and Domain Consensus; these features be modeled and integrated to evaluate the terms. Snow ball technique was used to acquire hyponymy relation, and HowNet-based method for extract general relation.

Others works like [17] present methods for semi-automatic class extraction using a database of Spanish verbs, diathesis alternations and syntactic-semantic schemes (ADESSE tool) [7], where the semantic extracted patterns are the classes. This methodology was applied in educational domain and replicated in financial domain in [16]; in both works, the class extraction was completed with the domain expert opinion. A method for class extraction using linguistic patterns and NLP metrics such as morphological labeling is presented in a recent research [15].

The researches which report ontologies manually constructed is based in the domain and evaluation, since the domain used for this proposal, mainly analyzes the works in pedagogical domain. In [22] an ontology for interaction between students and teachers for English language teaching is introduced. On the other hand, [13] proposes an ontology for the internet learning process. In both works is defined an ontology for each entity in the learning process, and the evaluation is conducted with a manual process supervised for domain experts. Other researchers are focused on online education ([21,3,4] and recently [12]) where ontologies are manually defined from XML resources available in the Internet, and the evaluation is a manual process too. An ontology created from CASE diagrams for on-line education is presented in [2]; its evaluation is addressed by experts in a manual process. In this researches, the focus is the construction step where the class are extracted manually.

There are works such as [20] focused on automatic learning; in this paper, an ontology based on the Internet of Things used in a classroom is created, considering the student intelligences. The ontology creation process from the courses information offered in advanced levels is explained in [1], where students can choose courses according with their academic background. Both works present the structure, information, and hierarchy of the classes in a manual way.

Some researches are focused in the corpus creation, but in different domains. [10] is focused on the creation of the relevant linguistic corpus written in Serbian language. The focus is the sentiment analysis of student generated contents on higher education. In [19] the problem of creating a reference corpus for the classification of news items in fine grained multi-label scenarios was analyzed. The authors propose a semiautomatic approach for creating a reference corpus that uses three auxiliary classification methods: Support Vector Machines, Nearest Neighbor Classifiers and another based on a dictionary.

3 Methodology

Figure 1 shows the steps proposed, where two parallel methods was worked: corpus and dictionary. For the corpus construction some academic researches was used with two principal characteristics: focused in social sciences (pedagogy) and contain papers written in Spanish language. Secondly, papers about the principal class was extracted and joined in a initial corpus.

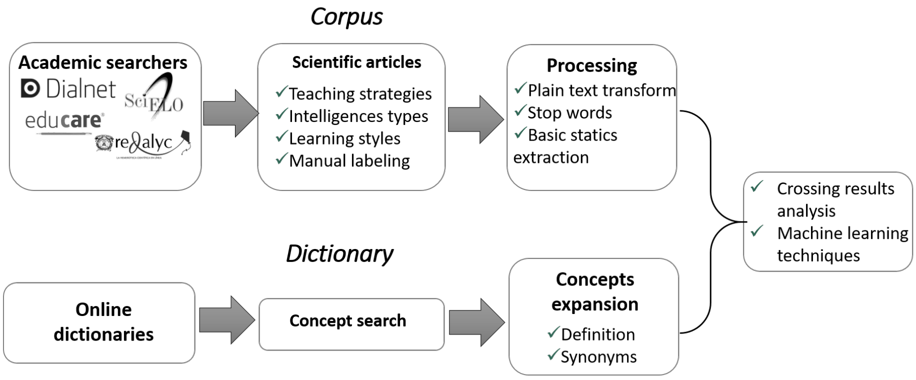

Fig. 1. Corpus built process.

In first step, four academic researches was choose:

- **Dialnet[1]:** Is is a project started in Rioja University. The site web collects and gives access to documents published in Spain in any language about Hispanic topics.
- **Scientific Electronic Library Online (SciElo)[2]:** It is a model for cooperative electronic publication of scientific journals in Internet. Its goal is to respond the scientific communication needs in developing countries (Latin America and Caribbean)
- **Educare Electronic Journal[3]:** Is a quarterly international electronic magazine of Costa Rica University. The journal diffuse the science production and develop the academic analysis in the education domain.
- **Redalyc[4]:** It is a online information system for scientific magazines in Latin America, the Caribbean, Spain and Portugal.

The dictionary was constructed using 5 principal resources: 2 pedagogical dictionaries, definition of Spanish Real Academy, encyclopedic dictionary (print edition) and online synonyms dictionary. The complete process are explained in section 5.

4 Corpus Description

Corpus represents a set of data or a collection of texts used for linguistic analysis and Natural Language Processing. It is composed of words that are not designed by linguists, but the authentic words that occur naturally in written or spoken language. From the aspect of corpus linguistics, corpus must meet two basic conditions: representative and balanced [18]. For this reason, the corpus is created using a reality sample; in this case, with published papers in the domain. Three principals topics was used for papers research:

1. **Learning styles**: Project the way of which a person learn. However, there exist alternatives about how is possible to learn concepts and processing information by humans. In different works, have been proposed several theories for describe the different types of learning. This work adopt as reference the Neuro-Linguistic Programming model (NLP), which implement a technique that allow to improve the communication level between instructors and students. Three styles were analyzed: visual, auditory and kinesthetic.
2. **Intelligences types**: An intelligence is the ability to solve problems, or to create products, that are valued within one or more cultural settings [8]. Humans have a capacities range and potentials that can be employed in productive ways (together or separately). This idea origined the multiple intelligences theory. The types of intelligence identified in [8]; linguistic, logical-mathematical, musical, bodily, spatial, interpersonal and intrapersonal.

[1] `https://dialnet.unirioja.es/`

[2] `http://www.scielo.org/php/index.php?lang=es`

[3] `http://www.revistas.una.ac.cr/index.php/EDUCARE/index`

[4] `http://www.redalyc.org/home.oa`

3. **Teaching strategies**: From teaching-learning methodologies described in literature, the Model-Eliciting Activities (MEA) will be analyzed, starting with the theory of mental processing [14]. These methodologies have been used from elementary education to professional education, with the propose to identify the way in which students learn concepts, and establish units, indicators and tools for conduct the investigation analysis.

Table A (appendix section) shows the papers selected for the corpus, including the publication year, first author and the magazine. Some thesis was used, but only the theoretical framework. The papers text was extracted and preprocessed in plain text. The result of this process was a corpus A with 21 instances, where each instance is a paper. A can be described such as $A = \{K, T, C\}$ where:

- K is a paper key and a numeric attribute $\{1...21\}$.
- T is the all paper text, including the title and abstract. in this texts, stops words, numbers and words with length less to 2 letters was deleted.
- C is the instance class, this is a nominal attribute according to the principal topic in the paper. $C = \{LearningStyle, IntelligenceType, LearningStrategy\}$. Each paper was manually labeled according its title; the corpus was balanced, thus exists 7 instances for each class.

Table 1 shows the vocabulary frequency in each class, this analysis was realized afterward initial preprocessing.

Table 1. Corpus Vocabulary.

Class	Words	Vocabulary
LearningStyle	13,228	3,528
LearningStrategy	19,264	4,450
IntelligenceType	24,609	6,162
Total	57,081	9,625

IntelligenceType class has more vocabulary, but this class contain 7 principal concepts, the, the difference between the class is justified. Analyzed the total; the class shared many words.

5 Dictionary Description

The construction process was similar to corpus construction. First a initial list of class and components were created manually, this list contains the subdivision of the principal class. The words were the follows:

actividad
auditivo
autoevaluación
corporal
espacial
aprendizaje
instrumentos
inteligencia
interpersonal
intrapersonal
kinestésico
lingüística
matemático
modelo
mental
musical
naturalista
prototipo
realidad
reutilización
transmisión
visual

Secondly, the concepts was searched in the resources mentioned in section 3. Then, a initial dictionary D_i of 22 concepts was created. D_i can be described such as $D_i = \{W, C_1, C_2, C_3, C_4, S\}$, where:

- W is the principal concept (initial list).
- C_1, C_2, C_3, C_4 are the definition in each of the dictionaries used. Some words, do not have definition in some resource. The texts was preprocessed deleting stops words, punctuation marks and words with length less to 3 letters.
- S a list of word synonyms. The list was preprocessed deleting the punctuation marks in plain text.

D_i was expanded using the words of definition and synonyms list for obtain a dictionary D_f. This dictionary is a simple list than only contain the concepts. Figure 2 shows an example with a generic word (no included in D_i).

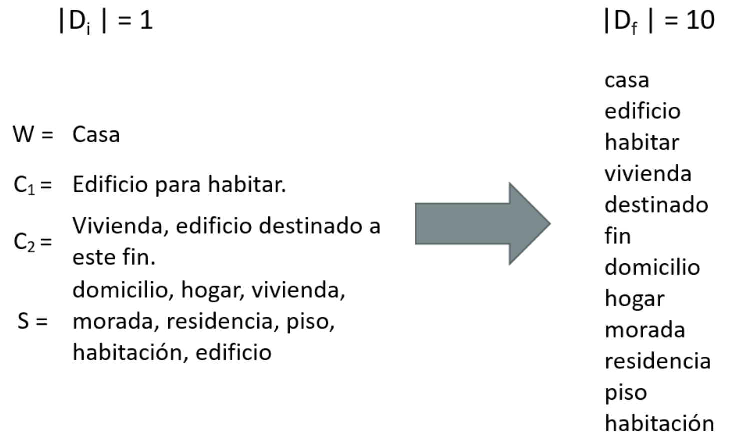
|D$_i$ | = 1
W = Casa
C$_1$ = Edificio para habitar.
C$_2$ = Vivienda, edificio destinado a este fin.
S = domicilio, hogar, vivienda, morada, residencia, piso, habitación, edificio
|D$_f$ | = 10
casa
edificio
habitar
vivienda
destinado
fin
domicilio
hogar
morada
residencia
piso
habitación

Fig. 2. Example of dictionary expansion.

Applied the expansion process, the list increased to 1,047 words. A final process was applied and the words with unitary frequency was deleted; then $\|D_f\| = 337$.

6 Preliminary Results

In this section, the result of crossing process was analyzed. Each of the D_f concepts was searched in A and the coincidences were counted. The number of matches are presented as follows:

LearningStyle	188
LearningStrategy	219
InteligenceType	239
Total Vocabulary	258

The IntelligenceType class has more words, but it is for the vocabulary size. In total, the number of coincidences is 258. The words of D_f that no appear in the corpus was deleted, thus $\|D_f\| = 258$. Finally, Table 2 shows the words of A with more frequency that appear in D_f

Table 2. Words in corpus with more frequency.

Word	Frequency
aprendizaje	895
estilos	420
inteligencia	397
inteligencias	272
educacion	256
actividades	243
proceso	227
ser	219
manera	210
asi	210
ensenanza	207

The words presented are related with the definition of the principal class, thus, is possible start with the process for the ontological learning. Also, the corpus was analyzed using supervised classification algorithms. Table 3 shows the accuracy for random forest and support vector machine algorithms using the vocabulary and the dictionary words like attributes. The results are better with the dictionary words with 81% of accuracy, thus, using the random forest algorithm, the corpus and the dictionary are stable with the main classes for applied the ontology learning process in the future.

7 Conclusions and Future Work

In this paper the construction process of lexical resources for pedagogical domain was presented. The quantitative analysis was added, and the preliminary results shows affinity between the resources (corpus and dictionary).

Table 3. Words in corpus with more frequency.

Algorithm	Vocabulary	Dictionary
Random forest	0.619	0.809
SMO	0.666	0.761

In future word, an analysis using a tagger in Spanish will be realized for to integrate words with similar sense. Then, it is necessary add to the dictionary hyperonyms for complete the semantic analysis. This research is the first step for applied the ontological learning in pedagogical domain, thus, the future word involve the ontology construction process using a semi supervised approach.

Acknowledgements. We would like to thank Vicerrectoria of Research and Graduate Studies, Benemérita Universidad Autónoma de Puebla for supporting this work through the project Model of Teaching-Learning Process applying Ontological Engineering.

References

1. Ameen, A., Khan, K.U.R., Rani, B.P.: Creation of ontology in education domain. In: 2012 IEEE Fourth International Conference on Technology for Education. pp. 237–238 (July 2012)
2. Bagiampou, M., Kameas, A.: A use case diagrams ontology that can be used as common reference for software engineering education. In: 2012 6th IEEE International Conference Intelligent Systems. pp. 035–040 (Sept 2012)
3. Dai, X., Li, X.: Study of learning source ontology modeling in remote education. In: 2010 International Conference on Multimedia Technology. pp. 1–4 (Oct 2010)
4. Du, L., Zheng, G., You, B., Bai, L., Zhang, X.: Research of online education ontology model. In: 2012 Fourth International Conference on Computational and Information Sciences. pp. 780–783 (Aug 2012)
5. El-Ansari, A., Beni-Hssane, A., Saadi, M.: A multiple ontologies based system for answering natural language questions. In: Rocha, Á., Serrhini, M., Felgueiras, C. (eds.) Europe and MENA Cooperation Advances in Information and Communication Technologies. pp. 177–186. Springer International Publishing, Cham (2017)
6. Fu, J., Jia, K., Xu, J.: Domain ontology learning for question answering system in network education. In: 2008 The 9th International Conference for Young Computer Scientists. pp. 2647–2652 (Nov 2008)
7. García-Miguel, J.M., Vaamonde, G., Domínguez, F.G.: Adesse, a database with syntactic and semantic annotation of a corpus of spanish. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
8. Gardner, H.: Estructuras de la Mente. Fondo de Cultura Económica (Sep 2001), `http://educreate.iacat.com/Maestros/Howard_Gardner_-_Estructuras_de_la_mente.pdf`

9. Grandbastien, M., Azouaou, F., Desmoulins, C., Faerber, R., Leclet, D., Quenu-Joiron, C.: Sharing an ontology in education: Lessons learnt from the oural project. In: Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007). pp. 694–698 (July 2007)
10. Grljević, O., Bošnjak, Z.: Development of serbian higher education corpus. In: Computational Intelligence and Informatics (CINTI), 2015 16th IEEE International Symposium on. pp. 177–181. IEEE (2015)
11. Guarino, N., Masolo, C., Vetere, G.: Ontoseek: content-based access to the web. IEEE Intelligent Systems and their Applications 14(3), 70–80 (May 1999)
12. Hssina, B., Bouikhalene, B., Merbouha, A.: An ontology to assess the performances of learners in an e-learning platform based on semantic web technology: Moodle case study. In: Europe and MENA Cooperation Advances in Information and Communication Technologies, pp. 103–112. Springer (2017)
13. Hu, J., Li, Z., Xu, B.: An approach of ontology based knowledge base construction for chinese k12 education. In: 2016 First International Conference on Multimedia and Image Processing (ICMIP). pp. 83–88 (June 2016)
14. Johnson-Laird, P.N.: Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press, Cambridge, MA, USA (1983)
15. Kang, Y.B., Haghighi, P.D., Burstein, F.: Cfinder: An intelligent key concept finder from text for ontology development. Expert Systems with Applications 41(9), 4494–4504 (2014), http://www.sciencedirect.com/science/article/pii/S0957417414000189
16. Ochoa, J.L., Hernández-Alcaraz, M.L., Almela, A., Valencia-García, R.: Learning semantic relations from spanish natural language documents in the financial domain. In: Proceedings of the 3rd International Conference on Computer Modeling and Simulation, held at Mumbai, India. Chengdu: Institute of Electrical and Electronics Engineers, Inc. pp. 104–108 (2011)
17. Ochoa Hernández, J.L.: Desarrollo de una metodología para la construcción automática de ontologías en español a partir de texto libre. Ph.D. thesis, Departamento de Ingeniería de la información y las comunicaciones. Universidad de Murcia (2011)
18. Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning. A Guide to Corpus Building for Applications. O Reilly Media, USA (2013)
19. Teixeira, J., Sarmento, L., Oliveira, E.: Semi-automatic creation of a reference news corpus for fine-grained multi-label scenarios. In: Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on. pp. 1–7. IEEE (2011)
20. Uskov, V., Pandey, A., Bakken, J.P., Margapuri, V.S.: Smart engineering education: The ontology of internet-of-things applications. In: 2016 IEEE Global Engineering Education Conference (EDUCON). pp. 476–481 (April 2016)
21. Wu, H.: Research of internet education system based on ontology. In: 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery. vol. 4, pp. 602–605 (Oct 2008)
22. Zhu, F., Fok, A.W.P., Ip, H.H.S., Cao, J.: Engonto: Integrated multiple english learning ontology for personalized education. In: 2008 International Conference on Computer Science and Software Engineering. vol. 5, pp. 210–213 (Dec 2008)

A Corpus description

Table 4. Papers list of initial corpus.

Author(Year)	Magazine Congress	Title
Morales, R. (2017)	Campus Virtuales	Inclusión de estilos de aprendizaje como estrategia didáctica aplicada en un AVA
Gómez, E. (2017)	Revista virtual Universidad Católica de Chile	Estilos de aprendizaje en universitarios, modalidad educación a distancia
Salas, J. (2014)	Revista Electrónica Educare	Estilos de aprendizaje en estudiantes de la Escuela de ciencias del Movimiento Humano y Calidad de Vida, Universidad Nacional Costa Rica
Bañuelos, A. (2014)	Esseñanza e Investigación en Psicología	Estilos de aprendizaje y su transformación a los largo de la trayectoria escolar
Cala, R. (2014)	Estilos de aprendizaje	Determinación de los estilos de aprendizaje de estudiantes de 1er curso de ing. Industrial y electrónica de la universidad técnica del Norte, Ibarra Ecuador
Tocci, A (2013)	Estilos de aprendizaje	Estilos de aprendizaje de los alumnos de ingeniería según la programación neuro linguïstica
Aragón, M. (2009)	CPU-e Revista de investigación Educativa	Diagnóstico de los estilos de aprendizaje en los estudiantes: Estrategia para elevar la calidad educativa
Álvarez, S. (2015)	TECTZAPIC Revista de divulgación científica y tecnológica	Actividades reveladoras del pensamiento que activan la reflexión matemática en un grupo de cálculo diferencial
Sánchez, G. (2015)	Proceedings of the Satellite conference of the International Association for Statistical Education	Implementación y transadaptación de actividades reveladoras del pensamiento (MEAS) en la enseñanza de estadística en el niuvel medio superior y a nivel superior. Un caso de estudio
Hernandez, F. (2013)	VII Congreso de Investigación, Innovación y Gestión Educativas	Actividades reveladoras del pensamiento: Una alternativa para desarrollar competencias matemáticas
Osorio, G. (2012)	Biblioteca Online ITESM (Tesis)	Las actividades reveladoras del pensamiento como estrategia de evaluación formativa en el área de matemáticas en el quinto grado de educación primaria
Alvarez, S. (2012)	Biblioteca Online ITESM (Tesis)	Actividades reveladores del pensamiento que activan el pensamiento matemático de los estudiantes dentro del proceso de las aplicaciones de la derivada que utilizan máximos y mínimos
Valdespino, E. (2011)	Biblioteca Online ITESM (Tesis)	Análisis de las respuesta de los estudiantes al trabajar una MEA con tablas de multiplicar en educación básica
Domínguez, A. (2009)	X Congreso Nacional de Investigación Educativa	Actividades reveladoras del pensamiento: más que una forma de aprendizaje activo
Garzón, A. (2016)	Revista de pedagogía	La i ntegración TIC-Inteligencias múltiples (IM): Una oportunidad de cambio en el proceso educativo
Barraza, R. (2016)	Actualidades investigativas en Educación	Rendimiento académico y autopercepción de inteligencias múltiples e inteligencia emocional en universitarios de primera generación
Mainieri, A. (2015)	Actualidades investigativas en Educación	Conocimientos teóricos y estrategias mtodológicas que emplean docentes de primer ciclo en la estimulación de las inteligencias múltiples
Inciarte, N. (2012)	Multiciencias	Inteligencias múltiples en la formación de investigadores
Juárez, J. (2010)	Investigación y Posgrado	Inteligencias Múltiples: Una innovación pedagógica para potenciar el proceso enseñanza aprendizqje
Paniagua, L. (2008)	Revista Electrónica Educare	La teoría de las inteligencias múltiples en la práctica docente en educación preescolar
Guzman, B. (2005)	Revista de investigación	Las inteligencias múltiples en el aula de clases