World Wide Web in the service of schooling: Semantic Web as a solution for language teaching in Cypriot secondary education*

Neofytou Chrystalla

Open University of Cyprus, Cyprus

chrystalla.neofytou@st.ouc.ac.cy

Abstract. This paper examines some suitability aspects of existing web search engines in relation to the content and the stated learning objectives of language teaching in Cypriot secondary education, focusing on the language course of the third high school grade (G9). The end goal is to put the internet in the service of schooling; specifically to categorize the results returned by the search engine into genres in order to facilitate user (teacher or student) in choosing the most appropriate texts for their learning purposes. The tools for categorizing texts are being sought in the field of Semantic Web technology, such as metadata, ontologies, software agents, and, the techniques in the fields of Natural Language Processing (NLP), Information Retrieval (IR), Information Extraction (IE) and Text Mining. The paper proposes the categorization of texts into six major genre categories according to their external (structural) and internal (linguistic, stylistic) characteristics. For the purpose of this research, the MeDa13 metadata model was designed on the basis of the standard metadata model Dublin Core, and the Textual Genres Ontology (TeGO) was developed for describing the concepts mentioned in genres. In this work, we present the theoretical background for the development of the proposed models (MeDa13 and TeGO), and also the methodological plan to achieve the research objective, which is the categorization of texts into genres considering the content and learning objectives for language teaching.

Keywords: Semantic Web (SW), Metadata, Ontologies, Natural Language Processing (NLP), Information Retrieval (IR), Information Extraction (IE), Text Mining, Greek Language Teaching, Cyprus Educational System.

* This is joint work with my PhD advisor, Dr Thanasis Hadzilacos, written and presented as a single author work for the purposes of the Doctoral Consortium of the Eighth European Conference on Technology Enhanced Learning, Paphos, 17 September 2013

1 Introduction

The internet as a source of information and a means of communication, either synchronous (chat) or asynchronous (email, blogs), has a dominant place in all fields of human everyday life, including education. The fact that searching and information gathering is the most common activity on the internet, brings up several questions regarding the real usefulness of the internet in school education and the suitability of the results returned by the existing classic web search engines in relation to the context and the learning objectives as set out in the curriculum of Cyprus Ministry of Education and Culture. Assuming that the internet is useful for the results returned to the user, this study seeks answers regarding the suitability of the results returned focusing on the learning purposes of language course of the third high school grade (G9). The frequent return of unsuitable results, as noted in the international literature, is related to the operation of existing web search engines, which are based on using keywords and conducting searches that are related more to the word spelling than with semantics [1]. The difficulty of automatic recognition of semantic content of the information stored in the Web, known as lack of semantics, orients the research interest of this study in the field of Semantic Web (SW).

The end goal is to design and develop a system that will distinguish textual genres, for example, a history text from a literary text, a text of the daily press from a sketch etc, and will organize internet data according to the textual characteristics of each genre. The first step is the definition of each (textual) genre and the clear wording of assessment criteria for categorizing texts. The long term vision is to build a search engine that returns the search results categorized into genres. This search engine will operate on the basis of specific assessment criteria for the texts, search the internet, filter all the data and sort them in the categories of genres. Therefore, the research looks for a tool that will 'help' the computer to process data, to understand the meaning e.g. of a historical text and the definition given for this genre, in order to filter the texts and subsequently to classify them into the genre categories that have been created. This means that the information stored in the computer should be semantically enriched. The research is oriented in the field of SW and seeks the tools with which the semantic enrichment of the information will be achieved, and, subsequently, the classification of internet data in the categories of genre created for the needs of this research. Additionally, the study looks for a tool that will select texts from all the available texts on the web and correctly register them in the appropriate categories.

This work proposes the automatic categorization of Greek texts that are available on the internet into genres, as presented and used in the school textbook for teaching Modern Greek language in the third high school grade (G9). To achieve this goal, an algorithm is developed and the relative techniques for categorizing texts are inquired in the fields of Natural Language Processing (NLP), Information Retrieval (IR), Information Extraction (IE) and Text Mining. The tools of SW technology, such as metadata and ontologies, are used to define the concepts or terms and the relationships between them, and to describe the related knowledge of a specific thematic domain, e.g. Medicine, Education etc., respectively. For the purposes of this study, a metadata model called MeDa13 is developed on the basis of the standard metadata

model Dublin Core, and the Textual Genre Ontology (TeGO) is formed in order to represent the knowledge which is relevant to genres.

2 Research Questions

The main research questions of this study are summarized as follows:

1. Does the operation mode of the internet and existing search engines meet the learning objectives of schooling, particularly for language teaching? Are the results returned from the existing search engines suitable for the content and the learning objectives of language teaching?
2. What would be the ideal scenario for the operation of the internet in relation to the educational needs? What would make an 'ideal' search engine? *(ideal = it would serve the learning objectives, return the most appropriate results in relation the query set by the user).*
3. Is the appropriate technology to 'place' the internet in the service of schooling available? Is SW the solution for ensuring the suitability of the results returned, always in connection with the content and the intended learning purposes? Which tools are available (metadata, ontologies, software agents etc.) and, which of them can be used to better organize the material on the internet in order to make it suitable for use in language teaching? Are the existing models suitable for the purposes of this work?
4. How should the material on the internet be organized? Which techniques for categorizing texts are available and which of them serve the objectives of this work?

3 Current knowledge of the problem domain

Search engines usually operate on the basis of keywords depending on the query entered by the user. The fact that these searches relate more to the spelling of the keyword than to semantics often results in unsuitable results. This is due to the construction and operation mode of the existing web search engines and their weakness to carry out the necessary semantic correlations between the query and the real objective of the user's search. Therefore, while the whole process of web search fascinates its users, at the same time, it is indifferent to computers while as machines fail to understand and interpret the information stored in them. The result is the creation of a communication gap between humans and machines related to human’s ability to read and interpret a word, phrase or sentence by associating it with the appropriate conceptual content resulting either in general or specific findings or reaching on reasonable inferences after the juxtaposition of two or more truthful sentences. Computers are unable to proceed to such automated correlations and (reasonable) inferences. For example, when humans read the sentence "Ο Μιχάλης είναι μεγαλύτερος από τον Αντρέα" (Michael is older than Andreas) easily conclude that "Ο Αντρέας είναι

μικρότερος από τον Μιχάλη" (Andreas is younger than Michael). The machine cannot understand the relative age of the two subjects, which is expressed using the comparative degree of the adjective "μεγάλος" (old) and, also to comprehend the syntax used to compare two objects or subjects in Greek language. Furthermore, the phenomenon of multiplicity, i.e. the use of one word to describe different objects, for example, the word "language" that can be interpreted as "anatomical organ", "communication tool", "fish species" etc, and the phenomenon of synonym, i.e. the existence of two or more words that describe the same object or situation but differ in style, performance or expressive significance, for example, the word "clever" and its synonyms "smart", " intelligent", "very clever" etc, make it even more difficult for the search engine to understand the question (query) set by the user. SW based on the creation of a common semantic basis (framework) allows the automation of these functions with minimum human intervention aiming to produce meaning and retrieve the most suitable information.

4 Existing solutions

The solution to overcome the structural and functional disadvantages of World Wide Web (WWW) seems to be SW, which is built upon the foundations of the existing web. SW, as the next step, structures (organizes) data correlating them with objects and entities of the real world, and, provides the necessary knowledge for any field of interest (e.g. Medicine, Education etc.). Using SW tools such as metadata and ontologies, computers are equipped with the knowledge of a specific subject area and the appropriate tools (data information, vocabularies, etc.) that enable them to 'read', process, interpret and understand the content of the information stored in them. SW bridges the communication gap between machines and humans based on the idea of having a commonly perceived, between computers and human beings, semantic framework of concepts and terms that refer to real objects and entities in the real world, expressing their mutual conceptual and semantic relations and representing the relevant knowledge. A semantic search engine returns the related to the search objective results after carrying out the essential conceptual correlations, linking concepts and objects with the most suitable semantic content.

Semantic Web is an initiative of the WWW Consortium (W3C) that was inspired by the creator of WWW, Tim Berners-Lee. Its goal is to structure information and to improve the current web, placing a semantic layer that allows machines to understand and process the (human) information effectively [2-3]. This web semantic enrichment is based on the vision of making the available information machine readable and understandable by equipping computers with the knowledge -in the form of dictionaries- so that they can understand the semantic content that concepts and terms carry. The basic principles of SW are to maintain the distributed web content, the representation and retrieval of information, the representation of concepts of various subject areas (e.g. Education), which is achieved by using ontologies and the existence of software agents [4]. The term 'ontology' is taken from philosophy and is used, according to

Aristotle, to describe the science of speech of being. In the field of computer science 'ontology' is understood as a divided and shared understanding of some domain that can be exchanged between people and systems applications. It is a standard (formal), categorical (explicit) specification of the distributed (shared) conceptual representation (conceptualization) [5]. It provides the required knowledge and vocabulary for the description (representation) of a specific field of interest (W3W). Software agents are special programs undertaken to 'look' on the internet and gather information, on behalf of the user, from various sources that have semantic content [6]. The description of the relationships between concepts (or terms) and objects (or entities) is achieved by using metadata that is information about online resources or other objects that are machine understandable [7]. It is structured information that describes, explains, locates or facilitates recovery, uses or manages an information resource.

Concerning the existing text categorization techniques, NLP moves in the field of linguistic analysis of texts at different levels. These are the morphological, lexical, syntactic, semantic, discourse and pragmatic level [8]. NLP adopts linguistic theories and examines their computational effectiveness based on linguistic data in order to understand natural language and resolve any ambiguities [9]. IR refers to the automatic retrieval of documents that are related to the query inserted by the user. It checks the representation and the relevance of a document and a query. IE is closely connected to the field of NLP, as it automatically extracts structured information from unstructured and/or semi-structured machine-readable documents, a process that concerns mostly human language texts by means of NLP [10]. Text Mining refers to the process of deriving information from text involving the process of structuring the input text [11]. Some of its tasks are text categorization, text clustering and concept/ entity extraction.

5 Preliminary ideas

1. Concentration of text genres to be included in the list of categories and formulation of general definitions for each genre. Recording the characteristics for each genre that will be used as assessment criteria, and grouping the texts in the following six major categories: Informative, Scientific, Literary, Artistic, Multimodal and Foreign_ Language.
2. Development of the algorithm for categorizing texts into genres and formulation of rules and axioms (assumptions) that lead to reasonable conclusions regarding the distinction of one genre from another (classification criteria).
3. Design and development of the metadata model MeDa13 based on the Dublin Core, for the enrichment of digital recordings (text) information on the (text) genre. MetaDa13 includes the following elements (13): Title, Creator, Subject, Publisher, Rights, Date, Date Modified, Source, Type, Description, Language, Data Writing, Authorial Intention, and, has two applications. The simple application is used to categorize texts into the six major categories of genre (e.g. Informative) and includes only five elements: type, description, language, data writing and authorial

intention. The composite application is used for more specified categorizing specific (sub-categorization, e.g. Journalistic) and includes all thirteen elements.
4. Design and development of Textual Genre Ontology (TeGO) for the description of concepts (terms) relating to genres (definitions and attributes) and their relations.

6 Proposed Approach and Research Methodology

The course of the research process in this paper is as follows:

1. Literature review to gather information on how World Wide Web and Semantic Web, and the web search engines operate (descriptive approach).
2. Conducting an experiment to search and gather information from the internet, for the purposes of an educational activity designed according to the content and learning objectives of a randomly selected topic from the school textbook for teaching Modern Greek language in the third high school grade (G9). The aim is to investigate the suitability of the results returned from the existing web search engines in relation to the content and the learning objectives of the language course (experimental approach).
3. Study of the curriculum of Cyprus Ministry of Education and Culture and the school textbook for teaching Modern Greek language in the third high school grade (G9). Wording of general conclusions concerning the suitability or unsuitability of the results returned from the experiment.
4. Study of the available Semantic Web tools (metadata, ontologies, software agents) and examination of their suitability for the purpose of this work.
5. Study of the existing techniques for categorizing texts in the fields of NLP, IR, IE and Text Mining, and search for the most appropriate for the needs of this study.
6. List all textual genres included in the school textbook and categorization into six categories created for the purposes of this study: Informative, Scientific, Literary, Artistic, Multimodal (texts using many modes i.e. language, sound, picture) and Foreign_ Language. Wording of general definitions and recording of the characteristics for each category and subcategory of genre. Categories and subcategories are defined according to the external and internal characteristics of each genre.
7. Gathering sampling (text) material from the internet in two phases: random and semi-directed selection of texts, using Google search engine and the keyword 'language'. Registration of the results (sample texts).
8. Categorizing a hundred selected results (texts) according to the general definitions and characteristics of genres (step 5).
9. Set two control groups consisted of ten primary school teachers and ten secondary school teachers respectively, to test the validity of definitions and assessment criteria by distributing questionnaires.
10. Gathering the results from the questionnaires and comparison of teachers' answers between them and with the results of the initial categorization (step 7). Wording of conclusions of the process.
11. Development of an algorithm on the basis of the assessment criteria for automated text categorization in the proposed categories of genres.

12. Design and development of the metadata model MeDa13, consisting of thirteen elements, on the basis of the standard metadata model Dublin Core, in order to describe the relationships between concepts (or terms) and objects (or entities).
13. Design and development of the Textual Genre Ontology (TeGO) to represent the knowledge related to the field "language teaching" and to describe the concepts relating to genres and their relationships.

7 Discussion of how the suggested solution is different, new, or better as compared to existing approaches to the problem

The originality of this work lies in the fact that it is carried around the axis of the content and the stated learning objectives for the language course for the third high school grade (G9), focusing on the textual material used in accordance to the lines of the curriculum of Cyprus Ministry of Education and Culture. The target group is teachers and students that use the internet for gathering information in the context of an educational activity and, the body of texts examined and categorized into genres is mainly written in the Greek language. Regarding the examination of the unsuitability of the results returned from the existing web search engines this study suggests the categorization of internet data into genres using Semantic Web tools, i.e. metadata and ontologies, in order to achieve the learning objectives of language teaching. The importance of this work lies in the proposal for the construction of a semantic search engine that will serve the purposes of language teaching and schooling, in general, as it will return search results after carrying out the necessary semantic associations and, finally, categorizing them into genres, underlining at the same time the need of multi-classification according to the external and internal characteristics of each text.

References

1. Taibi, D., Gentile, M., and Seta, L., 2005. A Semantic Search Engine for Learning Resources. Recent Research Developments in Learning Technologies.
2. Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. Scientific American. May, 2001.
3. Taibi, D., Gentile, M., and Seta, L., 2005. A Semantic Search Engine for Learning Resources. Recent Research Developments in Learning Technologies.
4. Kanellopoulos, D., 2012. The benefits of Semantic Web in e-business. Modern Technical Inspection. Issue Nov-Dec 2012. http://www.technicalreview.gr/index. php?option= com_content &task=view&id=684
5. Gruber, T. R. 1993. A translation approach to portable ontology specifications. In: Knowledge Acquisition. 5: 199–199.
6. Kanellopoulos, D., 2012. The benefits of Semantic Web in e-business. Modern Technical Inspection. Issue Nov-Dec 2012. http://www.technicalreview.gr/index. php?option= com_content &task=view&id=684
7. Berners-Lee, T., 1997. World-Wide Computer", Communications of the ACM, Vol. 40, No. 2, Feb. 1997.

8. Liddy, E., D., 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.
9. Lembesi, P., 2004. Introduction to Computational Linguistic and Historical Review. Presentation 2. http://hermis.di.uoa.gr/compling/Penelope_Intro.ppt
10. Wikipedia, the free encyclopedia. Information Extraction. Page last modified on 18 April 2013.
11. Wikipedia, the free encyclopedia. Text Mining. Page last modified on 20 July 2013.
12. Berners-Lee, T., 1998. Semantic Web Road map. September 1998. http://www.w3.org/DesignIssues/Semantic.html
13. Dublin Core. http://el.wikipedia.org/wiki/Dublin_ Core
14. Gaitanou, P. & Gergatsoulis, M., 2006. Ontology management: extended study on the main problems and presentation of existent ontology library systems. In Proceedings of the 15th Pan-Hellenic Academic Libraries Conference, pp. 136-150, Patra, Greece, 1-3 November, 2006. http://conference.lis.upatras.gr/files/2.04. FullText.pdf
15. Hadzilacos, Th., 2011. WWW search environment for K-12 Education, Seminar at the Open University of Cyprus. Nicosia, Cyprus 2011.
16. Kessler, B., Nunberg, G., and Schütze, H., 1997. Automatic detection of text genre. In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the conference, 7–12 July, Madrid, (pp. 32–38). [San Francisco, CA]: Morgan Kaufmann Publishers, 1997.
17. Mbalkizas, N., 2006. Google Search Engine. Presentation 3. In teacher training for the usage and implementation of ICT in teaching practice (Training Level II). 2006. http://users.sch.gr/nikbalki/epim_kse/files/Parousiaseis/Google_SearchMach ine.pdf
18. Pedagogical Institute of Cyprus. Teacher's book: Modern Greek Language, 3rd Grade. Cyprus, 2008.
19. W3C Semantic Web Activity http://lpis.csd.auth.gr/mtpx/ sw/