Usage of Ontologies for Multilingual Information Retrieval in the Semantic Web

Evgenia Vassilakaki
PhD student
Dept. Information and Communications,
Manchester Metropolitan University,
Manchester, UK
e.vassilakaki@mmu.ac.uk

1 Research Problem

The wide spread adoption of world wide web (www) and the passage to Semantic Web (SW), the increasing amount of information produced every day in all languages and the need to gain access to such a multicultural information have generated interest in the development of techniques for enabling multilingual Information Retrieval (IR). Multilingual IR (MLIR) is defined as the task of searching either a specific database or a set of databases, or the www for relevant information by using criteria in a chosen language (preferable in mother tongue) and retrieve all documents which match all the search criteria, regardless of the language of the document or the indexed language and present them in a unified list [1][2].

In this context, a variety of techniques have been introduced and are still under development aiming to cope with and overcome a series of issues regarding linguistic and semantic barriers mainly generated from the need to efficiently and effective retrieve and gain access to such a multilingual and at the same time multicultural information.

2 Related Work

The need to gain access to diverse pieces of information by equally diverse user groups from different cultural and linguistic backgrounds was early identified and a series of techniques have been since generated within the limits of SW [1]. These range from terminology mapping and language translation processes to multilingual thesaurus and ontologies. In [3] an overview of terminology mapping is provided as a way to improve the retrieval effectiveness for users and an opportunity to overcome problems associated with multilingualism. However, [3] admit that users still require accessing a number of different sources in order to find associated terms.

On the other hand, [4] proposes language translation process as a way to deal with MLIR. The use of a dictionary- based approach is adopted in order to translate the query into different languages. Then a term disambiguation technique is applied to select the best translation term, and a query expansion technique to enhance the queries' retrieval performance. On the downside, a series of variables that result to poor retrieval performance are identified such as the quality of the dictionary used, term ambiguity, phrase translation, untranslated terms such as acronyms or technical terms, lost or alternation of meaning.

Closer to successfully achieving MLIR is the development and use of multilingual thesaurus. In this context, [1] specify three key aspects that should encompass a modern thesaurus in order to contribute to the improvement of inter-cultural communication: a) they ought to be multilingual, b) they should be semantically structured and c) they should assist user comprehension and hence recall and precision in IR. [1] admit that barriers regarding complexity of natural languages and terminology exist resulting to inconsistences in one-to-one equivalences between terms of different languages. Thus, co-operation between classification and language processing specialists is proposed as a way towards achieving a new format of ”macrothesaurus” that would support international communication and knowledge transfer.

In the light of this, [5] presents ontologies as a tool that combines the advantages of natural language querying and the power of SW. An ontology is defined as ”a formal, explicit specification of a shared conceptualization” [6]. In [5] ontologies are adopted as the most appropriate mechanism for conducting MLIR because they are language independent, they are able to manage concept definitions and deal with linguistic barriers such as: the many meanings of the words; their dependence on context; the changing of the meaning over time, as well as space; the linguistic boundaries; the semantic definition over time etc. Furthermore, [7] explore the development and use of a multilingual ontology (MLO) in an attempt to overcome problems regarding query and document translation techniques in IR. The main advantages of this approach can be summarized to the need for no merging phase and no dependency on automatic translators.

In the context of SW, the need to give a well-defined meaning to information in order to enable computers and people to work in cooperation has generated discussion on developing new standards and technologies [8]. In particular, the development and usage of a MLO is proposed as the most appropriate and efficient way to reduce semantic ambiguity and heterogeneity and at the same time to increase interoperability and information integration and thereby the enhancement of the global, cross- cultural communication [9][10].

3 Contribution

This study aims to explore how the development and usage of MLO can enhance the efficiency and effectiveness of IR and improve the results. It also aims to present the benefits and drawbacks of the use of MLO in a semantic multicultural environment. Thereby, the following objectives will be pursued:

1. To form the Strategic Design for developing a MLO.
2. To record the stages and explore the difficulties and problems of developing a MLO.
3. To apply and use the MLO in a semantic multicultural environment such as the www.
4. To record and measure the direct outcome from the experimental use of the MLO.
5. To record and measure the experimental context and other intervening variables and determine whether these could affect the intended results.
6. To analyze and evaluate the direct output from the usage of the MLO.

In the context of this study, it is anticipated that the development and further usage of a MLO will contribute to a better and semantically management of multilingual information, the formulation of an intelligent tool for managing multilingual information, the improvement of the IR process in regard of multilingual information, the retrieval of relative information in different languages and the improvement of the multicultural communication in the context of SW.

4 Evaluation

This study is mainly explorative and the methodologies that will be used are Evaluation Research in conjunction with Literature Review. The critical review of the literature will be carried out in terms of setting the research into context and forming strategic decisions for successfully copying with issues regarding the development of MLO. In particular, guidelines and best practices will be identified and further adopted in an attempt to avoid drawbacks and problems already encountered by other researchers in the field of interest.

Evaluation Research is a methodology that enables the process of determining in a measurable way whether the intended results were produced by comparing them with the criteria of success established from the beginning [11]. For the purpose of this study, the indications of success will be the efficiency and effectiveness of IR and the specificity of the retrieved results succeeded by the use of a MLO applied in a semantic, multilingual environment. In this context, the evaluation research method will be applied in order to record and measure the direct outcome from the experimental use of MLO, as well as the intervening variables that could affect the expected results and compare the outcome with the initial indications of success.

Both Literature Review and Evaluation Research will contribute to the verification or not of the hypothesis made: if indeed the use of a MLO enhances the efficiency and effectiveness of the IR in the context of SW or not.

References

[1] Jorna, K., Davies, S.: Multilingual thesauri for the modern world: no ideal solution. Journal of Documentation **57**(2) (2001) 284–295

[2] Chen, A., Gey, F.: Multilingual information retrieval using machine translation, relevance feedback and decompounding. Information Retrieval **7**(1-2) (2004) 149

[3] McCulloch, E., Shiri, A., Nicholson, D.: Challenges and issues in terminology mapping: a digital library perspective. The Electronic Library **23**(6) (2005) 671

[4] Adriani, M.: Ambiguity problem in multilingual information retrieval. In: Cross-Language Information Retrieval and Evaluation: Workshop CLEF 2000, Lisbon, Portugal. (2001) 156

[5] Vertan, C.: Querying multilingual semantic web in natural language (2004)

[6] Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2) (1993) 199–220

[7] Guyot, J., Radhouani, S., Falquet, G.: Ontology- based multilingual information retrieval (2005)

[8] Berners-Lee, T., Shadbolt, N., Hall, W.: The semantic web revisited. IEEE Intelligent Systems **May/June 2006** (2006) 96–101

[9] Mayfield, J.: Ontologies and text retrieval. The Knowledge Engineering Review **17**(1) (2002) 71–75

[10] Kabel, S., de Hoog, R., Wielinga, B.J., Anjewierden, A.: The added value of task and ontology based markup for information retrieval. J. ASIST **55**(4) (2004) 348–362

[11] Babbie, E.: The practice of social research. Wadsworth Publishing Company, Belmont, CA. (1998) Methodologies.