Towards a Multilingual Ontology for Ontology-driven Content Mining in Social Web Sites

Marcírio Silveira Chaves[1] and Cássia Trojahn[2]

[1] Universidade Atlântica, Oeiras, Portugal
[2] INRIA & LIG, Grenoble, France

Abstract. Social Semantic Web aims at combining approaches and technologies from both Social and Semantic Web. While Social Web sites provide a rich source of unstructured information, what makes its automatic processing very limited, Semantic Web aims at giving a well-defined meaning to the Web information, facilitating its sharing and processing. Multilinguality is an emergent aspect to be considered in Social Semantic Web and its realization is highly dependent on the development of multilingual ontologies. This paper presents Hontology, a multilingual ontology for the hotel domain. Hontology has been proposed in the context of a framework for ontology-driven mining of Social Web sites content. Comments are annotated with concepts of Hontology, which are labeled in three different languages. This approach facilitates the task of comments mining, helping managers in their decision-making process.

1 Introduction

Social Web focuses on social interaction mainly through comments in social sites. Its rapid growth has created a huge unstructured and multilingual knowledge base, what essentially makes its automatic processing very limited. On the other hand, Semantic Web aims at giving a well-defined meaning to information on the Web, better enabling cooperation between software agents and people [2]. Ontologies are the key ingredients in the Semantic Web, providing a formalized way for representing knowledge of a domain.

Social Web and Semantic Web have been integrated into the called Social Semantic Web [3]. A motivating scenario is ontology-driven mining of comments from Social Web sites. For instance, hotels web sites contain a wealth data of users comments, which often help guests to decide whether making a reservation. Furthermore, hotel managers have been interested in mining comments for better exploring users (customers) knowledge. Once comments are annotated with ontologies, managers can exploit semantic search for supporting their analysis task and decision-making process. An emergent aspect in such a scenario involves to consider its multilingual content.

The realization of the Multilingual Social Semantic Web is highly dependent on the development of multilingual ontologies. Different approaches have been

proposed for dealing with ontology multilinguality [1, 6]. However, in practical, few multilingual domain ontologies are freely available. Only 2.5% of the ontologies in the OntoSelect[3] library is multilingual [13].

This paper presents the main ideas behind Hontology, a multilingual ontology for the hotel domain. Hontology has been manually created and its current version supports English, French and Portuguese languages. Each concept and property of Hontology are manually annotated with different labels in these three languages. Although for dealing with the huge source of knowledge at the web scale, automatic methods for creating and populating ontologies are required, Hontology can be seen as a starting point to these approaches. Hontology has been proposed in the context of a framework for annotating comments provided by users in social web sites [4].

This paper is structured as follows. Section 2 gives an overview of the framework in which Hontology is being proposed. Section 3 shows the details of Hontology and describes the methodology we have followed to develop it. A comparison between Hontology and related ontologies is presented in Section 4. Section 5 discusses the main approaches we are exploiting for extending Hontology. Section 6 discusses on other related work. Finally, Section 7 concludes the paper.

2 Multilingual Ontology Application

The inspiration for a framework for annotating comments from Social Web sites comes from the gathered needs in Customer Knowledge Management (CKM) research [8]. Such comments constitute new information sources to be integrated into CKM companies initiatives. CKM is the combination of Customer Relationship Management (CRM) and Knowledge Management (KM). CRM is a strategic approach concerned with creating improved shareholder value through the development of appropriate relationships with key customers and customer segments [12]. On the other hand, KM is the collection of processes that govern the creation, dissemination and leveraging of knowledge to fulfill organizational objectives [9]. CKM is an organizational strategy that aims at managing knowledge about the customer.

The hotel domain is strongly affected by the comments written in social sites. These comments often help guests to decide whether making a reservation. On the other hand, Hotel managers need tools to better explore customers knowledge from Social Web to support their decision-making task. To that extend, we have proposed a framework that aims at integrating knowledge from Social Web to support CKM (Figure 1). A multilingual ontology is the central element in this framework.

The entry point of the framework is the set of comments (raw text) from Social Web sites, e.g., "booking.com" or "realtravel.com". Some Social Web sites provide together with the comments, the user profile (e.g., family with young children or mature couples). These data (comments and profile) are then pre-

[3] http://olp.dfki.de/ontoselect

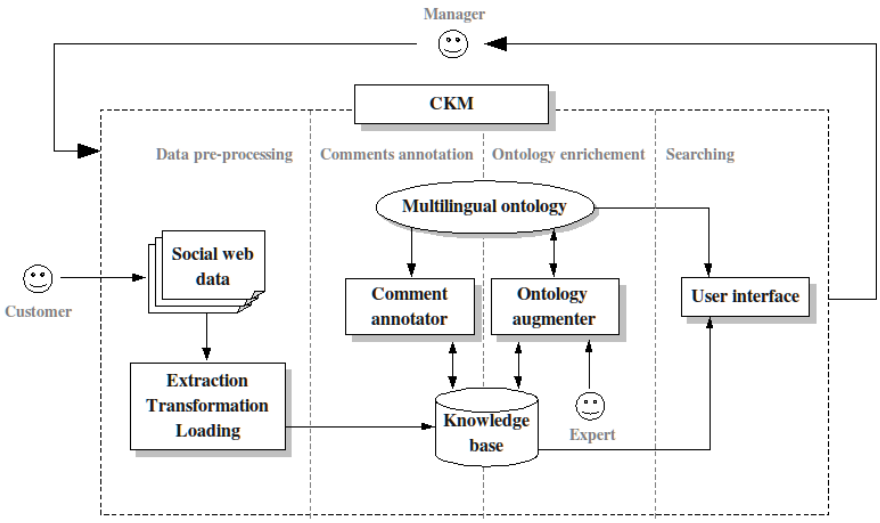

Fig. 1. A framework for CKM based on Social Semantic Web [4].

processed and stored into a *knowledge base* (KB). Three main processes are carried out in this pre-processing phase:

Content extraction deals with the selection of multiple information sources, i.e., different Social Web sites, and format, e.g., HTML, in which these information sources are provided;

Content transformation selects parts of the texts to be loaded, e.g., comment and its classification, and join data from multiple sources, e.g., join the comments of the profiles "family" with "family with young children", since information sources have different classifications to the profiles;

Loading classifies the comments, i.e., positive, negative or neutral, and store the classification into the KB. The classification task involves, basically, to identify adjectives, e.g., "good" and "satisfactory" or "not good" and "unsatisfactory" in comments in order to infer such classification.

Next, the comments are annotated with the concepts from the ontology, via the module *comment annotator*. This module receives as input the pre-processed comments and compares terms of these comments to labels of concepts in the ontology. Recognizing the language in which a comment is written is a precondition to annotate it. The comment is annotated with the corresponding ontology concept if the degree of similarity between them is above a specific threshold. In a first approach, we combine syntactic approaches, i.e., based on string similarity, such as string equality, sub-string and edit distance. Comments and their annotations are then stored into the KB.

Furthermore, Hontology can be augmented with new concepts, via the module *ontology augmenter*. This module is responsible for identifying potential new

concepts in comments, which are then filtered out and validated by a *user expert*. This identifying process considers terms correlation, rules (lexical patterns) and synonyms, as detailed in Section 5.

Finally, the *manager* can navigate within the concepts of the ontology and retrieve the comments annotated with the corresponding concepts. The advantage of using a multilingual ontology-driven navigation is two fold. On the one hand, once the manager has selected a concept (in English, for instance), all comments in all languages the ontology supports (including the synonyms of these concepts) are presented to the manager. This facility is not available in traditional search engines. On the other hand, it allows for the manager specializing and generalizing queries in an intuitive way, according to the ontology hierarchy. For instance, searching for the concept "Laundry", both sub-concepts "Laundry room" and "Laundry service" are included in the search as well as their synonyms ("pressing", for instance) and corresponding terms in different languages ("Blanchisserie" and "Lavanderia").

3 Hontology: A Multilingual Ontology for the Hotel Domain

3.1 Development Methodology

The development of Hontology has been carried out following seven main steps:

1. **Identify existing ontologies on related domains:** The first step was to search for ontologies on the hotel domain and related domains, such as tourism or traveling, in order to use them as a staring point for organizing the knowledge in our ontology.
2. **Select the main concepts and properties:** Based on the available related ontologies, we have filtered out the main concepts we have judged as interesting for considering in our ontology. These related ontologies are often built to be used on specific applications, using very specific concepts as well as they are monolingual. We compare Hontology with related ontologies in Section 4.
3. **Organize concepts and properties hierarchically into categories:** Hontology is neither a merge of existing ontologies nor an union of them. In the prior step, we have selected certain concepts and properties, which were manually re-structured into a hierarchy.
4. **Translate the ontology:** One of the main steps in the development of Hontology was to add, for each concept and property, different labels in different languages, including synonymous. This task was carried out by bilingual experts.
5. **Expand concepts and properties based on comments:** The process of expanding the ontology considers three main approaches, namely terms correlation, rules (lexical patterns) and synonyms, as detailed in Section 5.
6. **Translate the new concepts and properties:** The same as phase 4. For instance, the new concept "pillow" must have associated labels in French and Portuguese languages.

7. **Export the ontology in several formats:** Applications using Hontology explore different levels of formality. Some of them need to perform reasoning, while other just work with a flat list of concepts and properties. In order to satisfy these needs, we make Hontology available in OWL, RDF and XML.

3.2 Describing Hontology

Figure 2 presents the main concepts and properties of Hontology. For instance, lets consider the concept "Hotel Chain". In the right side of Figure 2, its associated labels in French and Portuguese are listed.

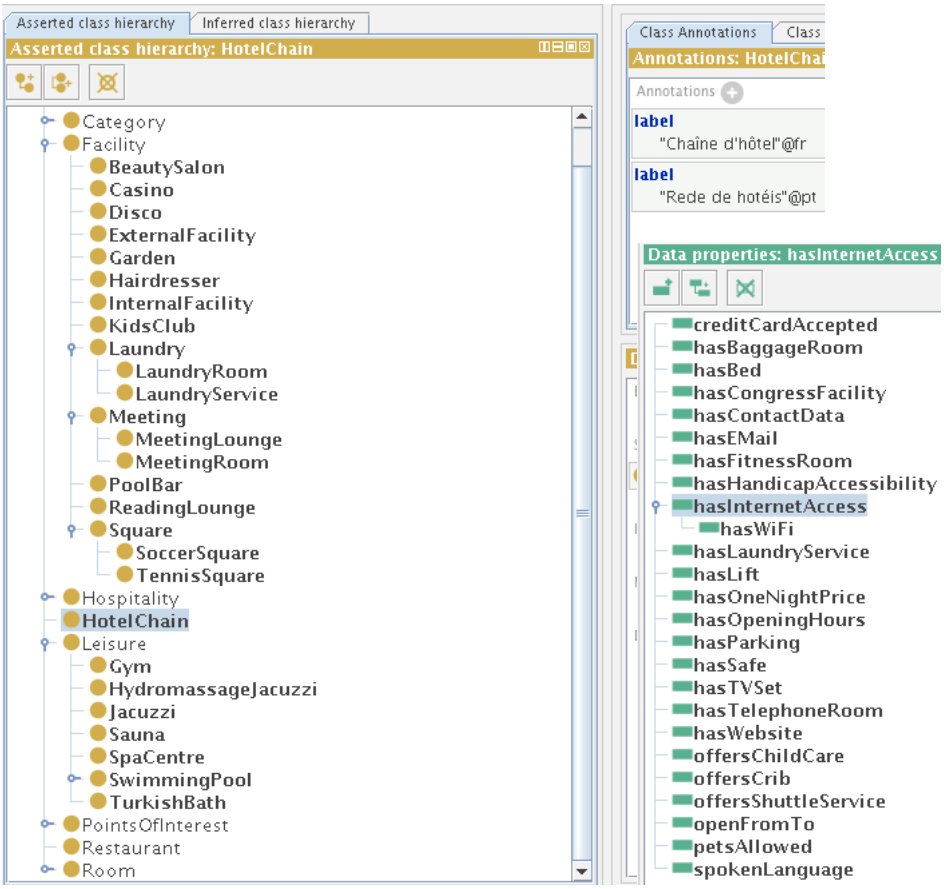

Fig. 2. The main concepts and the data properties of Hontology.

Hontology contains seven top concepts, which represent the corresponding sub-domains:

Category: contains all the types of categories into which a hotel can be classified, e.g., tourist, comfort and luxury.

Facility: includes the utility options offered by each hotel, e.g., beauty salon, kids club and pool bar.

Hospitality: contains the existing kinds of hotels, e.g., hostel, pension and motel.

Hotel: details the kind of hotels, e.g., bunker, cave and capsule.

Leisure: lists the leisure options, e.g., gym, jacuzzi, and sauna.

Points of interest: includes the main points, usually near to the hotel, which are most often mentioned in comments about the hotels, e.g., stadium, museum and monument.

Room: splits into Hostel Room and Hotel Room, which have different kinds and nomenclature for rooms.

In its current version, Hontology has 97 concepts, 9 object properties and 25 data properties. To the best of our knowledge, Hontology is the first multilingual ontology for the hotel domain. We have compared Hontology with related ontologies, as commented in the next section.

4 Comparing Hontology with Related Ontologies

Few ontologies for the hotel domain and related domains have been proposed, which cover different aspects in this domain. Furthermore, these ontologies are monolingual, what makes their use limited in the context of Multilingual Social Semantic Web. In this section, we compare Hontology with other relevant ontologies which have some relation with the hotel domain: Mondeca[4], HarmoNET[5] and Travel Itinerary[6]. Hontology is freely available at `mchaves.wikidot.com/hontology`. Ontologies describing concepts of hotels are often found as sub-ontologies of the tourism domain [5]. Table 4 presents a summary of the related ontologies.

Mondeca ontology is the largest one in number of concepts. However, it is neither public nor freely available, which restrict its usage. HarmoNET and Travel Itinerary are public and freely available. The first one describes accommodation and events concepts and the latter is used for representing traveling concepts. One limitation of these ontologies, as well as in Hontology, is that they do not contain instances associated to the schema. It is another aspect that restrict their reuse. In this sense, we are currently working on how populate Hontology with instances.

Hontology was built based on the concepts and properties of these existing ontologies. However, Hontology is not a merge of them. For instance, concepts like "Transport" and "Multimedia Item" from HarmoNET and "Flight" and "Meal" from Travel Itinerary are out of the scope of Hontology.

[4] `mondeca.com`
[5] `harmonet.org`
[6] `daml.org/ontologies/178`

Feature	Mondeca	HarmoNET	Travel Itinerary	Hontology
Multilingual	No	No	No	Yes
# concepts	1000	54	8	97
# properties	n.a.	166	24	34
# instances	0	0	0	0
Domain	Tourism	Tourism	Travel	Hotel
Use	Mondeca project	Accommodation/ event	n.a.	Hotel support decision
Public	No	Yes	Yes	Yes

Table 1. Comparison between Hontology and related ontologies.

This comparison evidences the lack of the multilingual ontologies in the hotel domain. Hontology is public and freely available for the research community and then can be used as a baseline for constructing new ontologies. This is an important point for promoting the development of new multilingual ontologies.

5 Extending and using Hontology

We are working on the improvement of Hontology, in the main directions commented below.

First, the module *ontology augmenter*, commented in Section 2, aims at enriching Hontology with relevant information from comments. This task exploits the following heuristics: term correlation, rules (lexical patterns) and synonyms. *Term correlation* considers potential terms mentioned in the comments, which are present in Hontology. For instance, in a comment containing the sentence "Rooms are comfortable, but pillows are very hard" the terms "pillow" (in the ontology) and "room" (not in the ontology) should be probably related through a property linking them in Hontology. Once the ontology is enriched with the term "pillow", a comment containing, for instance, only the sentence "Pillows are very hard" can be found under the concept "room".

Rules (or lexical patterns) consider that comments usually contain a set of common adjectives, e.g., "good", "cheap" and "soft". This approach uses lexical patterns and extract relevant terms which are preceding or succeeding the adjective, e.g., "Air-conditioned is loud", "Small bathroom".

Furthermore, *synonyms* are important elements that must be considered in the improvement of Hontology. They have already being considered in the process of adding labels to the concepts. However, this task can be extended with the help of dictionaries and lexical resources within an automatic process.

Second, we plan to work on multilingual ontology matching [7, 15, 16]. It is a primary problem to be solved, for instance, when integrating ontologies from different hotels. Our aim is to explore different kinds of labels (for instance, preferable and alternative labels), written in the same language, for helping in the matching task. Usually, multilingual matching tools use translations approaches or composition of alignments (a set of correspondences between two ontologies)

for dealing with the multilinguality in the ontologies to be matched (as in [16]). These approaches require external resources, such as translators and previously generated alignments, which are not always available for the languages being considered. Moreover, specially for languages deriving from the same root language, e.g., Latin, lexical and syntactic methods can be experimented in order to find potential alignments, as reported in [16]. Through alignments, we can link Hontology to other ontologies.

Third, we are working on extending Hontology for including labels of concepts in other languages such as Spanish and Italian. Two experts are currently working on this task. Moreover, from the existing concepts and properties in English, Portuguese and French, we intend to apply some techniques of ontology localization. We are considering to create a linguist information repository, such as in [13].

Finally, Hontology can be used as a multilingual resource to cross-language information retrieval. Cross-Language Evaluation Forum (CLEF)[7] has challenged multilingual systems to search in documents written in several languages. Queries and questions on hotel domain can be supported by Hontology, since the main concepts and properties are present in it. For instance, a query containing "hotels with jacuzzi" can be automatically translated to Portuguese and French with the support of Hontology.

6 Other Related Work

Multilinguality in ontologies has been exploited on different perspectives. First, tools for supporting automated inclusion of multilingual labels in ontologies have been proposed. Espinoza et al. describe a tool for automatically localizing ontologies, i.e., adapting an ontology to a concrete language and cultural community [6]. This tool translates labels in natural language and obtains a list of potential translations into the target language. The aim of this tool is reduced the human effort to localize ontologies.

Another approach involves to interface ontologies and lexical resources [14]. Ontologies and lexicons refer to different layers in the meaning representation. This comes from the fact that ontologies remain at the conceptual level of meaning representation, that is not *a priori* linked to any natural language. Dictionaries (and more generally lexical resources) are reference gateways to a language.

Although these efforts, no concrete ontologies have been made publicly available, as commented in Section 4. Our aim is to share with the community a multilingual ontology that can be extended by using the tools and approaches described above. Moreover, Hontology can be used as a baseline for evaluating this kind of tools.

[7] `www.clef-campaign.org`

7 Final Remarks

This paper has presented Hontology, a multilingual ontology for the hotel domain. Hontology gives support for annotating comments from Social Web sites in the context of a framework for Customer Knowledge Management.

In its preliminary version, Hontology has been manually created and supports three languages. Our main contribution is to make available for the community, a multilingual ontology that can be used as a baseline for many usages and applications in the context of the Multilingual Semantic Web, promoting its realization.

As future work, we plan to extend Hontology, in the following main directions: enrich Hontology by using potential terms from comments themselves; exploit Hontology in Multilingual Ontology Matching; include labels in other languages; explore issues related to ontology localization and internationalization. In addition, we plan to apply some machine-learning methods for the sentiment analysis on comments. The main idea is to classify a comment as "positive", "negative" or "neutral", for instance, what can hep hotel managers in their analysis. We plan to exploit method, such as SO-polarity (Subjective Objective) and PN-polarity (Positive-Negative), in order to determine the strong of comment PN-polarity, i.e., weakly positive, mildly positive, or strongly positive [10, 11]. Finally, we plan to populate Hontology with instances.

References

[1] Almeida, J. and Simes, A. (2006). T2O: Recycling Thesauri into a Multilingual Ontology. Fifth international conference on Language Resources and Evaluation, LREC 2006, Genova, Italy, May.

[2] Berners-Lee, T.; Hendler, J. and Lassila, O. (2001). The Semantic Web. A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities. Scientific American, volume 284, number 5, May. pp. 34-43.

[3] Breslin, J.; Passant, A. and Decker, S. (2010). The Social Semantic Web. IX, pp. 300, Hardcover, ISBN: 978-3-642-01171-9.

[4] Chaves, M.; Trojahn, C. and Pedron, C. (2011). A Framework for Customer Knowledge Management based on Social Semantic Web: A Hotel Sector Approach. In: Customer Relationship Management and the Social and Semantic Web: Enabling Cliens Conexus. Colomo-Palacios, Ricardo; Varajão, João and Soto-Acosta, Pedro (Eds.). Hershey, PA: IGI Global, 2011. (To appear).

[5] Choi, C.; Cho, M.; Kang, E. and Kim, P. (2006). Travel Ontology for Recommendation System based on Semantic Web. Advanced Communication Technology. ICACT 2006. The 8th International Conference. pp. 624–627.

[6] Espinoza, M.; Gómez-Pérez, A. and Mena, E. (2008). Enriching an ontology with multilingual information. Proc. of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications, Tenerife, Canary Islands, Spain, pages 333-347.

[7] Fu, B.; Brennan R. and O'Sullivan, D. (2010). Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web. Proceedings of the WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.

[8] Gebert, H.; Geib, M.; Kolbe L. and Brenner, W. (2003). Knowledge-enabled customer relationship management. Journal of Knowledge Management, 7(5), 107-123.
[9] Lee, C. and Yang, J. (2000). Knowledge value chain. Journal of Management Development, 19(9), 783-793.
[10] Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2, 1-2 (Jan. 2008), p. 135.
[11] Pang, B. and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of ACL, pp. 271–278.
[12] Payne, A. (2006). Handbook of CRM: Achieving Excellence in Customer Management, Burlington, MA, Butterworth Heinemann.
[13] Peters, W.; Montiel-Ponsoda, E.; de Cea, G. (2007). Localizing Ontologies in OWL. In: Proc. of the OntoLex07 Workshop at the 6th International Semantic Web Conference, Busan, South-Korea, November 11th.
[14] Prevot, L.; Borgo, S.; Oltramari, A. Interfacing Ontologies and Lexical Resources. Ontolex Workshop: Ontologies and Lexical Resources, October 15, 2005 Jeju Island, South Korea.
[15] Trojahn, C.; Quaresma, P.; Vieira, R. (2008). A Framework for Multilingual Ontology Mapping. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco. European Language Resources Association. pp. 1034–1037.
[16] Trojahn, C.; Quaresma, P.; Vieira, R. (2010). An API for Multi-lingual Ontology Matching. Proc. of the International Conference on Language Resources and Evaluation, LREC 2010, 17 - 23 May, Malte. European Language Resources Association.