Personalized Questions, Answers and Grammars: Aiding the Search for Relevant Web Information

Marta Gatius
Computer Science Department, Technical University of Catalonia,
Jordi Girona Salgado, 1-3, 08034 Barcelona, Spain
gatius@cs.upc.edu

Abstract

This work is about guiding the user web search by generating most relevant questions, answers and grammars from web documents. The proposed approach is based on the representation of the main domain concepts as a set of attributes and relating these attributes to the user models and to a syntactico-semantic taxonomy, that describes the general relationships between conceptual and linguistic knowledge. This taxonomy is used for both generating questions and answers and also for extracting data from the web. The data extracted from the web documents is represented as instances of the domain concepts. Questions, answers and grammars are generated from these instances.

1 Introduction

The large amount of data and services available on the web has increased the need of tools that may assist the different types of users when looking for information. The approach described in this paper to guide the user search consists of providing the most relevant data in a particular domain as a set of questions and their corresponding answers.

Presenting the main questions (and their answers) could be valuable in different types of scenarios, especially when the information to search is voluminous and/or when the user is looking for relevant data that has to be understood perfectly. For example, including most relevant questions and answers in the web description of academic courses could result useful for students, as described in the next sections.

The generation of personalized questions for a specific domain involves reasoning skills as well as domain and linguistic knowledge. To reduce the human effort needed in this process, this work proposes a general organization of the conceptual and linguistic knowledge involved, thus limiting the specific data that has to be incorporated for a new domain. In this proposal, the main domain concepts are described by a set of attributes and those attributes are related to user models and to a syntactico-semantic taxonomy, which represents the general relationships between conceptual and linguistic knowledge. This taxonomy, described in a previous work (Gatius, 2013), was defined following (Bateman et al, 1994). It is used for generating questions, answers and grammar and also for extracting data from the web.

This work is focused on the generation of questions and answers from (semi)structured web documents describing particular cases of general concepts (i.e., university courses and types of foods). Information from these documents can be automatically extracted and represented as instances of the general concepts, previously described by the expert. Questions, answers and grammars can be automatically generated from the resulting instances.

The next section gives an overview of the approach proposed together with its adaptation in several languages (English, Catalan and Spanish) for two different domains: university courses and cultural events. The Section 3 describes the implementation and evaluation done. Finally, related work and discussion is given in the last section

2 Approach Overview

The approach proposed to generate personalized questions, answers and grammars from web documents in a particular domain is based in a separated representation of the different types of

Proceedings of The 10th International Natural Language Generation conference, pages 203–207,
Santiago de Compostela, Spain, September 4-7 2017.

knowledge involved. This approach consists of the following five steps:

1. Representing the most relevant domain concepts as a set of attributes.
2. Relating the attributes to the taxonomy.
3. Relating the attributes to the users groups.
4. Extracting the web data.
5. Generating the personalized questions and answers (or grammars).

The first three steps, studied in a previous work (Gatius, 2015), have to be done by a human expert. First, the main concepts have to be defined by a set of attributes and these attributes by *facets*, which describe details such as the cardinality.

In a second step, the conceptual attributes have to be associated with the classes in the syntactico-semantic taxonomy, defined in previous work. These classes are associated with the linguistic structures involved in the questions and answers about the conceptual attributes. They can be easily adapted to new languages.

Figure 1 shows a partial description of the concepts involved in this scenario: **Course** and **Exam**. The course description is represented by a set of attributes. Each of the course exam is represented by an attribute, which value is an instance of the concept **Exam**. As can be seen in the Figure, the conceptual attributes have been associated with the corresponding syntactico-semantic classes. For example, the attributes **code** and **content** are related to the class **of**, corresponding to general descriptions and there are others attributes related to its subclasses, describing more precise information: **of_quantity** (i.e., the **credits**, **assessment** and **weight**), **of_time**, **of_date** and **of_place**.

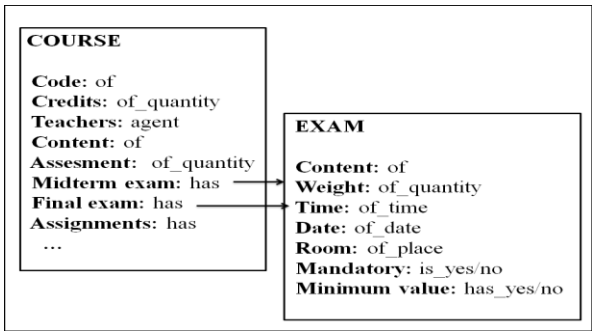

Figure 1: The main concepts describing a course

User models can also be incorporated by classifying users in groups and associating each conceptual attribute with the group interested on it. In several scenarios, stereotypes could also be related to different values of the attribute. For example, in the nutrition domain, the values of the attributes describing the number of calories and physical activity needed daily are different for each group (women, men and children).

In the academic scenario two user groups can be distinguished: students and teachers. The attributes describing the course are considered relevant for the two groups, except for the attribute **Code**, only interesting for teachers.

The two last steps proposed can be done automatically, although human supervision of the data obtained is needed. First, from the web documents selected, the appropriate data is extracted and represented as instances of the domain concepts. The values of the conceptual attributes are obtained using general rules defined for that purpose. Finally, the questions and answers (or grammars) are generated from the conceptual instances.

2.1 Personalized Questions and Answers

Most university web sites include clear and detailed descriptions about their courses. However, frequently, students ask teachers about this information, especially that related to the exams. For this reason, including relevant questions and their answers in the course description could help.

The extraction of the data in this domain requires a limited human effort, because the descriptions of university courses usually include similar content and are presented in a (semi)structured form. Several web documents from different faculties in the same university have been analyzed.

The web description of the courses analyzed, is placed in separated documents, with different formats: The particular data related to the exams (date, time and room) is presented by tables while more general information is in textual form.

The data related to a particular course is extracted from the web documents and represented as instances of the concepts **Course** and **Exam**. For this purpose, domain independent rules that use the facets describing the attributes (type, related terms and cardinality) are used. The data extracted is represented as the values of the instance attributes.

The general rule for obtaining the value of an attribute from a textual document is: "*If the attribute related terms (or synonyms) are found, then extract the **context** words that correspond to the type of the attribute*".

In this rule, **context** is a variable that indicates the maximum number of words before or after the attribute terms that have to be considered. Its value is obtained by analyzing the domain documents.

A condition to this general rule has been added to extract all possible values of the attributes, considering its cardinality.

Using this rule, the data describing the final exam of a particular course is obtained from the document giving general data and represented as the instance shown in Figure 2 (which belongs to concept **Exam** in Figure 1).

FINAL EXAM
Content: all units
Weight : 40%
Date: June
Mandatory: Yes
Minimun value: No

Figure 2: Representation of data extracted as an instance

1. **Which is the content of the final exam?**
   The content of the final exam is all units.
2. **How much is the final exam worth?**
   The final exam is worth 40% (of the final mark).
3. **When is the final exam?**
   The final exam is on June.
4. **At what time is the final exam?**
   The place will be detailed in the web course.
5. **Where is the final exam?**
   The place will be detailed in the web course.
6. **Is the final exam compulsory?**
   Yes, it is.
7. **Is there a minimum grade of the final exam?**
   No, there is not.

Figure 3: Generated questions and answers about the exam

More specific details about the exam have to be obtained from a separated document, where the date, time and room for the exams of several courses in the faculty are presented in a table. For this type of document a new rule is used:

If one of the course identifiers is found in a row then extract the next words in the row that correspond to the type of the attribute.

Figure 3 shows examples of the generated questions and answers obtained from the instance **Final Exam** in Figure 2.

2.2 Generating Personalized Grammars

Language interfaces have also been used to assist the user when accessing the web. They can incorporate domain-restricted grammars to help the user about the contents and the terms to be used to build the query, as can be seen in Figure 4. Those semantic grammars can also be generated following the approach proposed. The processing of the resulting query is simple, because the language to be considered is limited, i.e., the user will describe time by selecting one of the forms in the screen.

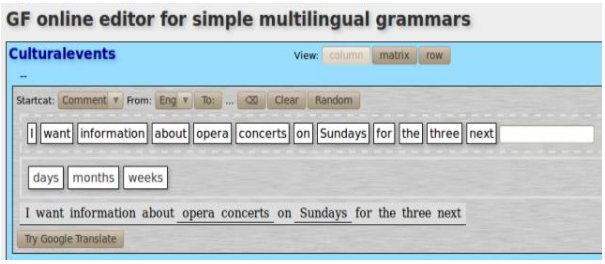

Figure 4: Guiding the user to build the query to the service

Figure 4 shows an example of how the user is guided to build a query to a web service giving information about the cultural events in a particular city. In this scenario, the grammar used has been generated from the concept **Event**, described by a set of attributes that correspond to the parameters of the service: **title**, **type**, **place**, **time** and **audience**. The same concept could be adapted for many of the web services about cultural activities. Two different user groups are distinguished, considering the value of the attribute **audience**, if they are interested on activities for adults or for children.

The interface shown in Figure 4 has been generated by the Grammatical Framework (GF, www.gramaticalframework.org), from the grammar written in the GF formalism, although other formalisms and environments could also be used.

3 Implementation and Evaluation

The web documents are first analyzed and classified in two groups considering their structure. Domain and language independent rules to extract the relevant data from these two types of documents have been defined and implemented in *C* language.

To automatically generate the personalized linguistic resources from the domain concepts, a *Prolog* program has also been developed. *Prolog* is

an appropriate language because its unification mechanism facilitates the association of general conceptual categories with features indicating additional information: stereotype, language and syntactic details (such as gender, number and tense).

Average punctuation for the 3 questions	Group 1 (0..10)	Group 2 (0..10)
1. Do you think the questions and answers about the course assessment have helped you to solve doubts?	8.10	8.43
2.Do you think is useful to incorporate questions and answers about the course assessment in the course web page?	9.14	8.78
3.Do you think you are going to consult again those questions?	7.8	7.74

Table 1: Results of the questionnaire

The questions and answers related to the exams of a particular course on introduction to programming were generated in three languages (English, Catalan and Spanish) and included in the course web page. In order to evaluate their usability, the students of two different degrees were asked to complete, anonymously, an online questionnaire, included in the same web page. There were 26 students in the Group 1, enrolled in the Bachelor's Degree of *Aerospace Vehicle Engineering* and 27 in Group 2, in the Bachelor's Degree of *Industrial Technology Engineering*. Table 1 shows the questions and their results, rating scales are from 0 to 10, 0 strongly disagree, 10 strongly agree. This result indicates that students think the generated questions and answers are useful: 8.4 over 10 in Group 1 and 8.12 in Group 2. Similar results were obtained from students in the same degrees, in an informal evaluation, done the previous semester.

4 Related Work and Discussion

The generation of questions and answers has focused many research works in different areas, such as educational (Wyse and Piwek, 2009) and conversational systems (Varges et al., 2006), (Okoye et al., 2011).

There are different techniques that can be used for generation, based on rules (Mazidi and Tarau, 2009) and/or statistical methods (Jin and Le, 2016). Those techniques can be adapted to textual documents and/or to structured data (Duma and Klein, 2013). In the first case, the generation process is usually done by applying rules to the trees obtained from the syntactic analysis (Nouri et al., 2011), although there are also works that use the resulting semantic structure (Kuyten et al., 2012) and other use both (Heilman, 2011).

Generation from structured data has been studied for years in language interfaces, which usually obtain the system inquiries and responses from application specifications and domain-restricted bases. Domain knowledge representation has been incorporated into a considerable number of relevant dialogue systems (Guzzoni et al., 2006; Sonntag et al., 2007), because they facilitate the adaptation of knowledge to different domains, languages, user types and modes of communication. Additionally, they provide synonyms, hyponyms and hyperonyms terms to improve the query.

There is an increasing interest in the combination of language and user model techniques to obtain personalized linguistic resources (Brusilovsky and Millán, 2007; Milosavljevic and Oberlander, 1998; Stock et al., 2007; Han et al., 2014).

This article describes an approach to guide the user about the web contents based on the generation of personalized questions, answers and grammars from web documents, because they could result useful in different scenarios, as the students opinion on the questions generated about course exams (shown in Table 1) indicates.

This work proposes an organization of the different type of knowledge involved (conceptual, linguistic and about the user) that minimizes the human effort needed for a new domain and/or a new language, by separating the general facts that can be reused across domains from those more specific. The representation is based on relating the set of attributes describing main domain concepts to the user models and to a taxonomy representing general relationships between conceptual and linguistic knowledge. The linguistic information associated with the taxonomy classes is used for both generating questions and answers and grammars and also for extracting data from the web.

Future work could include the study of the adaptation of the set of rules developed to extract the data from the web documents to new domains.

Acknowledgments

This work has been partially supported by the FreeLing project http://nlp.cs.upc.edu/freeling.

References

John Bateman and Bernardo Magnini and Fabio Rinaldi. The generalized *{Italian, German, English}* upper model, In Proceedings of the ECAI94 Workshop: Comparison of Implemented Ontologies, Amsterdam, 1994.

Peter Brusilovsky and Eva Millan. 2007. User models for adaptive hypermedia and adaptive educational systems. The Adaptive Web, 4321:3–53.

Daniel Duma and Ewan Klein. Generating natural language from linked data: Unsupervised template extraction. In the proceedings of the 10th International Conference on Computational Semantics, pages 83–94, Potsdam, Germany, 2013.

Marta Gatius. Adaptive Generation of Multilingual Questions and Answers from Web Content. In: 26th International Workshop on Database and Expert Systems Applications, Valencia 2015.

Marta Gatius. Improving Knowledge Representation to Speed up the Generation of Grammars for a Multilingual Web Assitant. In Fernandez Raquel AND Isard Amy (ed.). Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue semdial, December, 2013.

Didier Guzzoni, Charles Baur and Adam Cheyer. Active : A unified platform for building intelligent web interaction assistants. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops), pp. 417-420.

Xiwu Han, Somayajulu Sripada, Kit Macleod and Antonio Ioris. Latent User Models for Online River Information Tailoring. In the proceedings of the 7th International Natural Language Generation conference, 2014.

Michael Heilman. 2011. Automatic Factual Question Generation from Text. Ph.D. thesis, Carnegie Mellon University.

Yiping Jin and Phu Le. Selecting Domain-Specific Concepts for Question Generation with Lightly-Supervised Method. In the proceedings of the 9th International Natural Language Generation conference, pages 133–142, Edinburgh, UK, September 5-8 2016.

Pascal Kuyten, Timothy Bickmore, Svetlana Stoyanchev, Paul Piwek, Helmut Prendinger, and Mitsuru Ishizuka. 2012. Fully Automated Generation of Question-Answer Pairs for Scripted Virtual Instruction, volume 7502 of LNAI. Springer-Verlag.

Karen Mazidi and Paul Tarau. Infusing NLU into Automatic Question Generation . In the proceedings of the 9th International Natural Language Generation conference, 2016.

Maria Milosavljevic and Jon Oberlander. Dynamic hypertext catalogues: helping users to help themselves. In: HYPERTEXT, ACM, 1998.

Elnaz Nouri, Ron Artstein, Anton Leuski, and David Traum. 2011. Augmenting conversational characters with generated question-answer pairs. In the proceedings of the AAAI Fall Symposium.

Ifeyinwa Okoye, Jalal Mahmud, Tessa Lau, and Julian Cerruti. 2011. Find this for me: mobile information retrieval on the open web. In Proceedings of the 16th international conference on Intelligent user interfaces (IUI '11). ACM, New York, NY, USA, 3-12. DOI: https://doi.org/10.1145/1943403.1943407.

Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf, Norbert Pfleger, Massimo Romanelli, Norbert Reithinger, (2007), "SmartWeb Handheld — Multimodal Interaction with Ontological Knowledge Bases and semantic web services", Proceedings of the ICMI 2006 and IJCAI 2007 international conference on Artificial intelligence for human computing Pages 272-295.

Oliviero Stock, Massimo Zancanaro, Paolo Busetta, Charles Callaway, Antonio Krüger, Michael Kruppa, Tsvi Kuflik, Elena Not, and Cesare Rocchi. 2007. Adaptive, intelligent presentation of information for the museum visitor in PEACH. User Modeling and User-Adapted Interaction 17, 3 (July 2007), 257-304.

Sebastian Varges, Fuliang Weng, and Heather Pon-Barry. 2009. Interactive question answering and constraint relaxation in spoken dialogue systems. Natural Language Engineering.

Wyse, Brendan and Piwek, Paul (2009). Generating questions from OpenLearn study units. In AIED 2009 Workshop Proceedings Volume 1. The 2nd Workshop on Question Generation, 6-9 July 2009, Birghton, UK.