Adaptive Generation of Multilingual Questions and Answers from Web Content

Marta Gatius
Department of Computer Science
Technical University of Catalonia
Barcelona, Spain

Helping the user when searching for web information

- There are already many web crawlers designed to present a set of web documents containing information related to the user's query.
  - However, this does not always satisfy the user's information goal.

- Although the presentation of a precise answer, below the document level could be more useful, it is yet an unsolved problem.

Helping the User When Searching for Web Information

- The adaptation of conventional information retrieval and extraction techniques to the web faces new challenges:
  - The huge volume of documents
  - Data in different formats and languages
  - Hyperlinked information.

- Several techniques from several areas have been used to solve these challenges:

  use of semantic knowledge, information-seeking dialogues, incorporation of user models

The Generation of Personalized Questions and Answers

- Presentation of most relevant information in a specific domain
- Presentation of precise information (instead of complete documents)
- Related to the interactive approach to information retrieval, but does not require so much effort.

- This work studies how human efforts could be reduced by combining techniques from several disciplines to generate questions and answers:
  - **Semantic knowledge**, **user models**, **natural language generation** and **wrappers**

Semantic Models

- Different semantic models can be used:
  - Formal: database models, frames, ontologies
  - Informal: sets of attributes

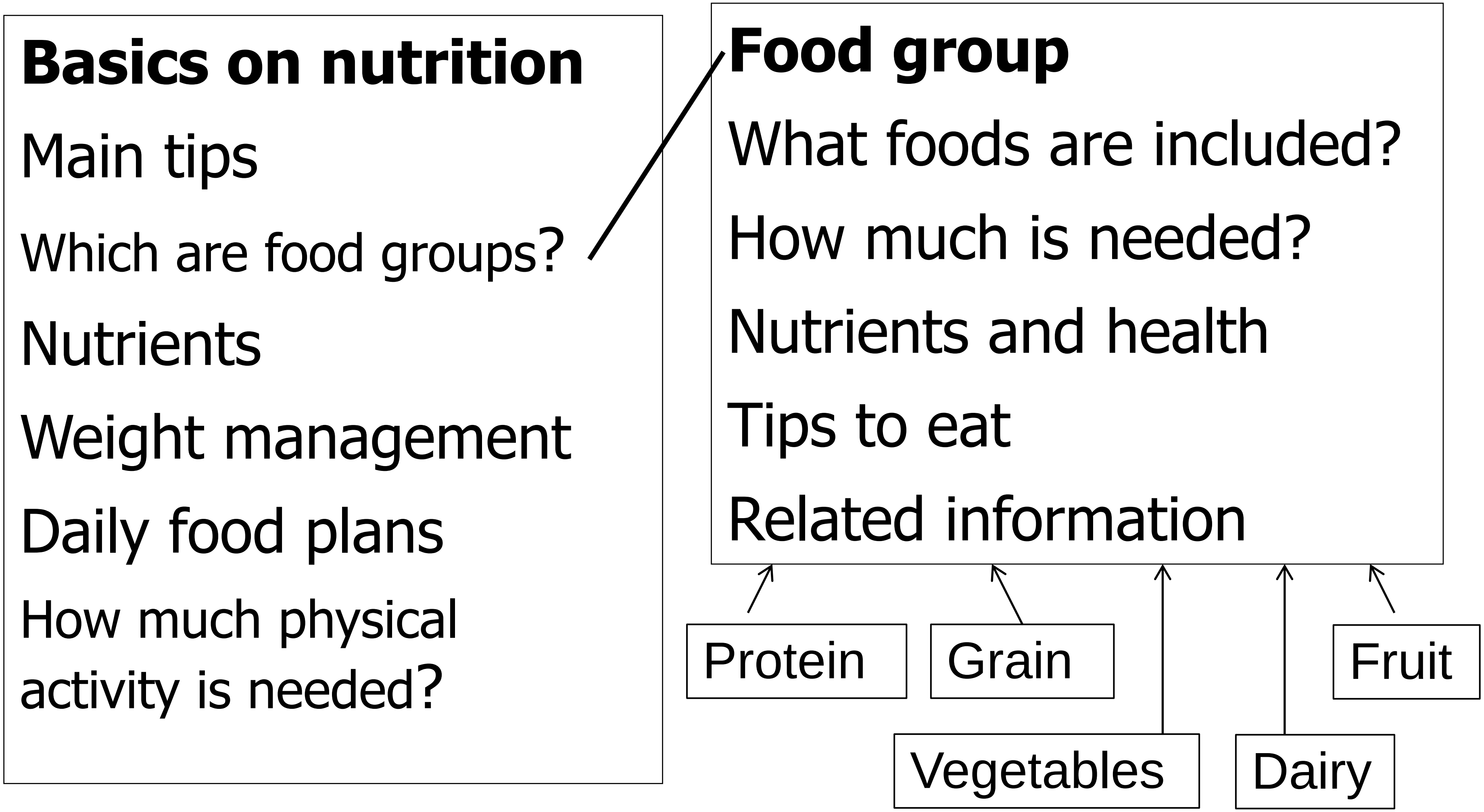
- The use of semantic models representing domain entities facilitates:
  - Generation of domain-restricted sentences
  - Multilinguality
  - Multimodality

Example of a Simple Concept
Academic Domain

Course
Complete name
Code
Academic year
Degree
Credit number
Teaching language
Coordinator Teacher
Other teachers
…

Example of a Complex Domain

Nutrition Domain



Basics on nutrition

Main tips

Which are food groups?

Nutrients

Weight management

Daily food plans

How much physical activity is needed?

Food group

What foods are included?

How much is needed?

Nutrients and health

Tips to eat

Related information

Protein

Grain

Vegetables

Dairy

Fruit

Ontologies

Advantages

- They provide a flexible way of representation
- Appropriate for integrating different sources
- Formal relations between entities (i*s-a, instance, part-of*) and preconditions
- Organization of knowledge in classes/subclasses
- Support under/over specifications

Disadvantages

- Expensive to build
- Difficult to reuse for different types of applications

Appropriate for complex domains

User Models

- Appropriate for adapting contents and applications for many users

- Have been used since the 80's:
  - User interfaces, e-learning, web contents

- Focused in a particular user feature:
  - Background (i.e. Profession), knowledge, interests, goals, individual traits

The User Background

- Distinction of groups (stereotypes)
  - Easy
    - Provided by the user
    - Simple to work with
  - Useful
    - Different groups - different contents
- Appropriate for specific domains
  - Academic domain: Teacher, Student
  - Medical domain : Professional, Patient

The User Background

- The background of a user group is represented as a layout over the set of **attributes** that represent the domain concepts.
- It consist of a **subset of the attributes** describing the domain concepts
- The attributes relevant for **more than one stereotype**
  - May have different values for each stereotype

Academic Domain

Two stereotypes: *Teacher and Student*

The attributes relevant for both have the same value

Course

Complete name ***Teacher***

Code ***Teacher***

Academic year ***Teacher/Student***

Degree ***Teacher/Student***

Credit number: ***Teacher/Student***

Teaching language ***Student***

Coordinator Teacher ***Teacher/Student***

Other teachers ***Teacher/Student***

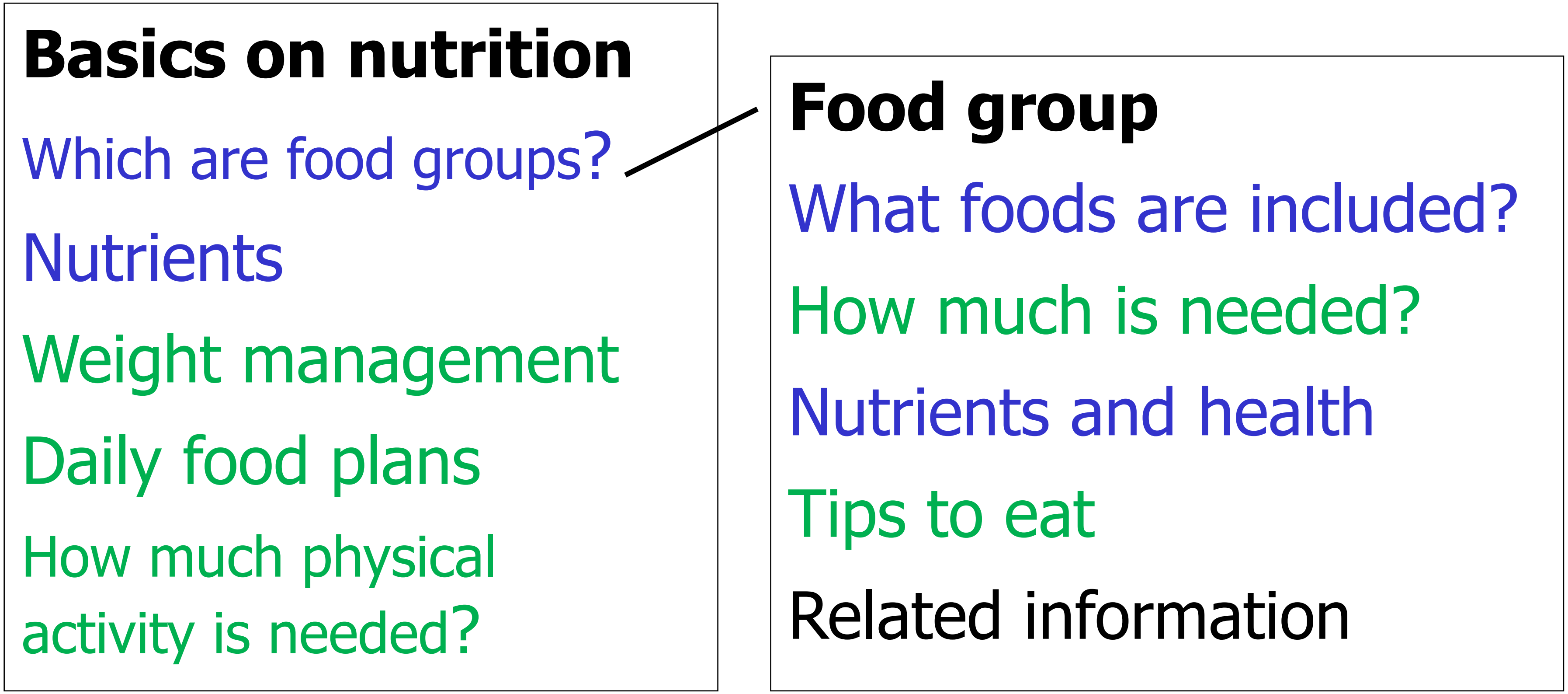
…

Nutrition Domain

Several stereotypes: considering age, sex and other (desease, profession)

Most attributes relevant for all

- Same value for all types
- Different value



Basics on nutrition

Which are food groups?

Nutrients

Weight management

Daily food plans

How much physical activity is needed?

Food group

What foods are included?

How much is needed?

Nutrients and health

Tips to eat

Related information

Natural Language Generation

The use of a syntactico-semantic taxonomy

- It relates the attributes describing the domain concepts to the linguistic structures needed for their expression
- Each attribute class is associated with several patterns to express questions and answers
- It supports several languages:

  Spanish, English and Catalan
- It facilitates the semi-automatically generation of language

The Natural Language Generation

Course

Complete name ***Teacher*** ***of_name***

Code ***Teacher*** ***of_code***

Academic year ***T/S*** ***of_year***

Degree ***T/S*** ***belong_to***

Credit number: ***T/S*** ***of_quantity***

Teaching language ***Student*** ***of_language***

Coordinator Teacher ***T/S*** ***who_does***

Other teachers ***T/S*** ***who_does***

…

Examples of Generated Questions and Answers for Students

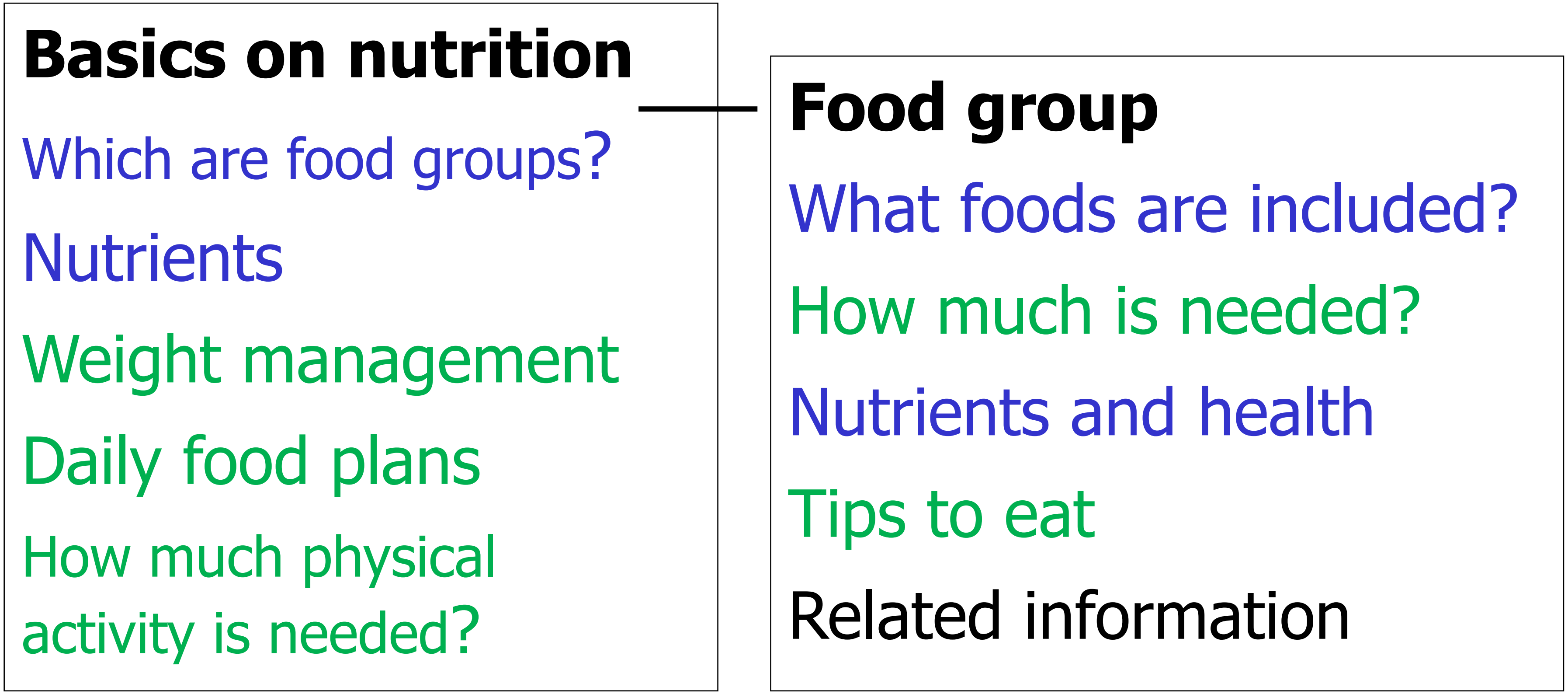
Q1. **How many credits does the course has?**
The course has 6 credits

Q2 **Who is the coordinator teacher of the course?**
The coordinator teacher of the course is Dr. John Smith

Q3 **Which is the teaching language of the course?**
The teaching languages of the course are Catalan and Spanish

Nutrition Domain

- Most web sites present information as questions and highlighted sentences
- Only selecting the most relevant information for each user is needed



Basics on nutrition

Which are food groups?

Nutrients

Weight management

Daily food plans

How much physical activity is needed?

Food group

What foods are included?

How much is needed?

Nutrients and health

Tips to eat

Related information

Wrappers Generation

- Wrapper systems use delimiter-based methods (such as HTML tags) to extract specific information in a particular web-document.

- A system of wrappers, previously developed, has been integrated to reduce the effort of obtaining web information that changes frequently, such as dates of exams.
  - It can obtain text from tables and captions from pictures

The Adaptive Generation of Questions and Answers

A separated and declarative representation of the different types of knowledge involved :

- **Conceptual**
  - Domain concepts are represented by a set of **attributes**
- **User model**- User background
  - A layout over conceptual knowledge
- **Linguistic** - Domain Lexicon

  - General relations between conceptual and linguistic knowledge

The Adaptive Generation of Questions and Answers

- The general process in a particular domain would be:

1. Study of the needs of user’s types.
2. Describing the domain concepts as a set of attributes
3. Relating each conceptual attribute with the specific type of user and the syntactico-semantic category.

- This general process can be adapted to new domains, knowledge models and languages.
- It has been implemented for generating simple questions and answeres about course descriptions.
  - **Prolog** language was used because of the unification mechanism

The Adaptive Generation of Questions and Answers

- The human effort needed would depend on:
  - The concepts involved
  - The organization of the data (in several interlinked documents, several formats)
  - The number of user types.

- In case web sites are already organized as questions and key sentences the human intervention needed can be reduced to select the most relevant of them for each stereotype, as well as the associated value
  - Wrappers can used for this purpose

Conclusion

- This presentation is about the generation of personalized questions and answers from web contents to satisfy users' information goal, either it is obtaining a specific data, casually exploring or getting a general idea.

- This work is based on a modular organization of the different types of knowledge involved: Conceptual concepts, user background and linguistic knowledge

Conclusion

The organization of knowledge is based on a description of the domain concepts by a set of attributes. And each of those conceptual attributes is associated with

- the user background.
- the syntactico-semantic class, that defines the linguistic structures for questions and answers

Conclusion

- The representation of knowledge in separated and declarative sources facilitates its adaptation to different domains, user models and languages.

- However, the cost of adapting this proposal to a complex domain can be high.

- In several web sites this process can be simplified
    - If data is already organized as questions
    - If information related to each user type is clearly presented

Future work

- The evaluation of questions and answers generated from courses descriptions.
  - They could be incorporated in the web sites describing the course to evaluate if they help students find the information.

- The study of how most relevant domain concepts can be automatically extracted from web documents.

Acknowledgment

We thank the Spanish project SKATER

THANK YOU FOR YOUR ATTENTION

The Syntactico-semantic Taxonomy

- The basic classes correspond to grammatical roles: Participants, being, possession, relationships and descriptions (i.e. Name, quantities, time and place)
- Subclasses have been defined considering further linguistic details
- Classes have been associated with patterns for expressing questions and answers
- Attributes in the class **of_quantity** can be realized with general patterns
- *How many <attribute-name> has the <concept-name> ?*
- *How many credits has the course ?*