EnviroInfo 2007 (Warschau)
Environmental Informatics and Systems Research
 ISBN: 978-3-8322-6397-3

Towards New Generation Environmental Information Services

Leo Wanner[1]

Abstract

Current environmental information services still too often offer "raw" pollution figures in terms of tables and graphics – possibly accompanied by pointers to general background information on the nature of the pollution and its relevance to human health. However, we must be aware that, on the one hand, non-specialists are not able to assess and interpret raw figures, and, on the other hand, certain kinds of pollution (e.g., air pollution) have a highly differentiated impact on different individuals. Therefore, new generation environmental information services must make recourse to techniques of Artificial Intelligence and Computational Linguistics (such as interpretation and reasoning, natural language generation, etc.) to provide adequate information. In this paper, we focus, first of all, on the language aspect. We elaborate first on the requirements with respect to environmental information services and assess then the MARQUIS service in the light of these requirements.

1. Introduction

When talking about environmental information services, we still too often think of tables and graphics, with a possible reference to general background information on the nature of, for instance, the individual air pollutant substances and their relevance to human health. However, we must be aware that

(i) non-specialists (in particular, general public, but also decision makers in politics and industry) are not able to assess and interpret "raw" pollutant concentrations in order to draw proper conclusions for their own behaviour, respectively decisions;
(ii) the kind and amount of the information relevant to the information addressee tends to depend on his/her health, prior knowledge and cultural characteristics; often, it is essential to go beyond the mere pollutant concentrations;
(iii) the mode in which the information is to be communicated depends on the nature of the data and, again, on the characteristics of the addressee; while tables and graphics may be adequate for specialists, text presentation is certainly more adequate for general public.

New generation environmental information services must take these aspects into account in order to be adequate, i.e., to satisfy increasingly demanding users and to provide the national and regional authorities a means to comply with the European environmental legislation. More precisely, they must, firstly, consider that the information cannot be uniform for all its users. As in the case of any information service, different types of potential addressees exist – with all consequences this implies with respect to need studies, user typology construction, etc. Secondly, they must incorporate models for assessment and interpretation of raw environmental data. Such models can be conceived either as rule-based *expert system shells* or as *machine learning* applications. Thirdly they must integrate information production programs (natural language text generators, graphic generators, and table generators). "Off-the-shelf" production programs are available. However, they still require a considerable amount of work in that the document planning techniques must be adapted to the domain in question, linguistic resources (lexica and grammars) must be extended to cover all linguistic constructions to be used while rendering the information to

[1] Institucio Catalana de Recerca i Estudis avançats and Pompeu Fabra University, Passeig de Circumval.lació, 8, 08003 Barcelona, Spain, e-mail: leo.wanner@upf.edu

the addressee, etc. In what follows, we elaborate on selected aspects of next generation environmental information services, focusing in particular on the role of Natural Language Processing (NLP). For illustration, we draw upon the *Multimodal Air Quality Information Service for General Public* (MARQUIS) (Wagner, 2007).[2]

2. The challenge of environmental information production

Dynamic information services on which we focus here are of great importance for all environmental topics that change periodically sufficiently fast to make manual authoring tedious, repetitive and time-consuming (see also Section 3). The topics involve, e.g., floods, forest fires, daily meteorological conditions (in particular weather), and air quality – including pollutant substances and pollen. Especially the latter two call for an automation of the process of information production. In principle, the user can be interested in the state of affairs at the time of the inquiry, the state of affairs in the past, or in a forecast of the state of affairs. While for meteorological forecasting rather accurate models are already available, air quality forecasts still pose a challenge. However, the following threefold challenge is common to both topics: (a) how to derive content from raw data (in other words: how to turn raw numeric time-series into information); (b) how to tailor the content to the user, and (c) how to tailor the information presentation to the user and to the content that is to be communicated.

For all three aspects of this challenge, a detailed user typology is a prerequisite. Additional primary topics are assessment and interpretation of raw data in the light of the user profiles and contextual settings and information generation.

2.1 User modelling

User modelling is a well-known problem in Artificial Intelligence (e.g., Paris, 1993; Zukerman/Litman, 2001). In the case of environmental information production, user modelling requires, on the one hand, empirical studies that identify the interests of the different types of users directly or indirectly affected by the environmental conditions, differences in cultural, language, etc. background and personal preferences, and, on the other hand, medical competence that allows for the identification of environmental conditions that affect specific types of users. The result must be a dedicated user model typology.

A user model is thus indicative for the content that is of relevance to a given user and for the way this content is to be rendered into tables, graphics and wording. Given the highly diverse profiles of the potential users and thus the impracticality to capture all of their details in a necessarily restricted user typology, the option of an individualization of system predefined user models by the user him-/herself seems very desirable. Such individualization may concern in particular the kind content a user is interested in. User model individualization facilitates thus the realization of a *collaborative* content selection strategy, which assumes that the service assesses the content with respect to its relevance to the given addressee, but leaves the final decision which information is to be delivered to the addressee him-/herself. This is in contrast to the *authoritative* content selection strategy, which assumes that the service is the only authority to decide upon the content to be offered to a given user. However, in certain contexts, the communication of specific information is essential. That is, a service should realize a combination of the authoritative and collaborative strategies – with the authoritative strategy having priority over the collaborative one. For instance, if in the case of the communication of air quality information an alarm threshold of a pollutant

[2] The development of MARQUIS has been supported by the EC in the framework of the eContent program (EDC-11258).

concentration is reached, the associated health warnings must be issued regardless of whether this option has been chosen by the user or not.[3]

2.2 Assessment and interpretation of raw data

The raw environmental data as measured by monitoring networks usually distributed over the geographical regions of interest require an assessment and interpretation with respect to: (i) expected state of affairs (e.g., level of air pollution, concentration of individual air pollution substances, weather conditions, etc.) in short term future, i.e., forecasting; (ii) relevance to the user of the state of affairs at the time of reporting and in short term future. While forecast provision is generally recognized as a necessary service for the population,[4] the elaboration on the state of affairs in terms the user can understand and the assessment of its relevance to the specific types of users from the corresponding user typology is still far from standard. The consequences could be observed when two years ago elderly citizens especially in France severely suffered from a heat wave due to insufficient information policies. Still, meteorological conditions are better assessed and communicated than air quality conditions; consider, for instance, the "Bio Weather" column in many newspapers, weather TV channels, etc. The assessment and interpretation of air quality conditions is in general much poorer. To be adequate, it must address at least the following topics (see also Nicklaß this volume) to derive content from the raw data:

- projection of concrete concentrations of pollutant substances onto European and region-specific air quality indices,
- health relevance of the observed or forecasted pollution for specific groups of users,
- significant air pollution changes over a certain time period and exceedance of European and national thresholds,
- justification of air quality conditions with meteorological conditions exploiting the correlations derived from the assessment models or provided by specialists.

Especially health relevance assessment is important since air pollution is to be viewed from distinct angles for healthy people, people with respiratory diseases, people with weak heart conditions, joggers, etc.

2.3 Generation of information

Once the content has been derived from the raw data, it needs to be rendered into the appropriate form the addressees can comprehend. It is to be borne in mind that environmental information is communicated via different communication channels (printed media, mobile services such as SMS, MMS, WAP, telephone, TV, internet, email, etc.) using text, graphics and language as modi. The language mode is either written (in the case of the majority of the channels) or spoken (in the case of a telephone and a TV service). For meteorological information, the use of spoken language is much more common than for air quality information.

The generation and presentation of information from the content obtained by means of assessing the raw data consists of four stages: (1) selection of the content that is relevant to the user that solicited information; (2) planning of the presentation of the information in especially the textual mode; (3) actual generation of the information in terms of graphics, tables and text; (4) planning of the layout of the pre-

[3] So far, in instruction and report generation, usually the authoritative content selection strategy has been implemented. Hypertext-based generation naturally favours the collaborative strategy (O'Donnell, 2001).

[4] Obviously, this does not mean that forecasting does not present any challenges. As already mentioned, especially air quality forecast models are still very much a topic of research. However, we do not delve into this topic in what follows.

sentation in case of written information. Bouayad-Agha/Wanner and Bohnet (both in this volume) elaborate on stages (1) and (2) and (3), respectively. Given the primary relevance of natural language processing (NLP) techniques for the aspect of the production of environmental information, we focus on it in the next section – before in Section 4 the air quality information service implemented in the MARQUIS-project is evaluated.

3. The role of NLP in environmental information services

In the contextual setting of environmental information services, NLP is first of all Natural Language Text Generation (NLG). Speech recognition and language analysis come into play in connection with a telephone based service. However, we will ignore both tasks here for the sake of simplicity and focus on NLG.

3.1 Why textual environmental information?

In general, the use of NLG for the production of information is considered to make sense only if the information domain in question fulfils certain criteria (Reiter and Dale, 2000). These criteria imply that

- the content changes periodically at a certain pace (or, at least often enough), which makes it necessary to rewrite the offered information each time the content changes,
- the text genre of the domain is monotonous and repetitive rather than creative and is thus for human writers more of a burden than literary work,
- the variation and complexity of linguistic constructions common in the information on the domain can be handled by a machine.

The environmental domain fully meets these criteria. Thus, the monitored environmental data (and thus the content derived from these data) potentially changes with each measurement (i.e., possibly as often as every 30 minutes); the assessment and interpretation of the monitored data (i.e., the content) may even differ between different users. The weather and air quality bulletins are tedious to write since they do not require any creativity from the writers. And, finally, both the variation and complexity of the linguistic constructions and the vocabulary of the environmental domain are limited.

But it is not only these three standard criteria which call for the use of NLG in environmental information services. Thus, as already argued above, environmental (and especially air quality) data need to be explained, tables and graphics need to be commented upon, and health related warnings need to be issued. In this context, it is important that the language in which the information is rendered be the language of the user. In other words, multilingual NLG is required.

3.2 Why "full-fledged" text generation?

Given that the need for automatic text generation is beyond any doubt, the way this generation is realized must be clarified. For some restricted applications, so-called "template-based" sentence generation suffices. A sentence template is a predefined, i.e., "canned", sentence with empty slots that are to be filled with content that is modified dynamically. Consider, for illustration, the following template:

1. *The ozone concentration* <development> <time-period>.

where the slot '<development>' can be filled, for instance, with one of the elements from the following verbal set: {*fell*, *rose*, r*emained stable*} and the slot <time-period>' can be filled with an element from the set {*early morning*, *morning*, *noon*, *afternoon*, *evening*, *night*}. Template-based sentence generation is straightforward and fast to implement. However, it is suitable only for very restricted scenarios in which a

limited number of sentence patterns suffices and different slot filler elements do not require a modification of the templates in which they are used. Thus, the AutoText UIS-service (Bohnet et al., 2001), which delivers information on one single air pollutant, namely ozone, in the region of Baden-Württemberg, Germany uses templates. However, as the authors of AutoText UIS report, already with a focus on ozone, template generation reaches its limits. That is, if one wants to cover all relevant pollutants of a region (see also Section 4 below) or provide a detailed weather report, more flexible generation techniques are needed. Such techniques involve two major processing stages: (1) the content and discourse structure planning stage, (2) the generation stage proper. In the literature, such techniques are often referred to as "full-fledged" text generation.

During the content and discourse structure planning stage (cf. also Bouayad-Agha/Wanner this volume), the content that is to be communicated to the user in question, the order in which the individual content chunks are to be communicated and the rhetorical strategy underlying this communication are determined. The result of this stage is a *text plan* that is further processed by the generation stage proper.

The generation stage proper consists in the mapping (or transformation – depending on the linguistic theory on which generation is based) of the content elements included in the text plan onto surface (cf. also Bohnet this volume). The mapping/transformation may contain several intermediate steps – again, depending on the linguistic theory. It requires the availability of generation grammars and lexica. The generation grammars specify how exactly content elements are to be mapped onto linguistic structures, and, if needed, more abstract linguistic structures are mapped onto more surface-oriented linguistic structures. The lexica contain the lexical information on each lexical unit (= word in one specific sense) and possibly also the mapping between conceptual and/or semantic units and lexical units.

Obviously, full-fledged generation is considerably more costly than template-based generation. However, its flexibility in the generation of coherent and cohesive textual information justifies the costs.

4. MARQUIS: A Next Generation Service?

The MARQUIS-service (Wanner 2007 and the contributions in this volume) has been built as a user-tailored multilingual multimodal air quality information service. In this section, we want to explore to what extent MARQUIS meets the requirements imposed on advanced environmental information services.

MARQUIS provides information for five selected European regions with respect to all relevant air pollutant substances.

4.1 User Modelling in MARQUIS

MARQUIS uses a user typology that has been established as a result of a thorough empirical study carried out within all regions covered by MARQUIS;[5] cf. Figure 1. Each user type is assigned a default user profile which specifies the content elements to be rendered and the modi (text, graphic or table) in which each content element is to be rendered. The default profiles can be individualized by the user to a certain extent in that she/he can choose to receive more detailed or more general information than foreseen by the default in the preferred presentation mode (see Bouayad-Agha/Wanner this volume for more details).

[5] About 100 individuals with different backgrounds were interviewed in each region.

Consultant/AQ-Technician:

1. Medical professional
   - 2.1 Asthma specialist
   - 2.2 Specialist for other respiratory diseases
   - 2.3 Specialist for non-respiratory diseases
2. Patient
   - 3.1 Asthma patient
   - 3.2 Patient- Other respiratory diseases
3. General Public
   - 4.1 Printed Media Reader
   - 4.2 Television Viewer
   - 4.3 Mobile Phone User
   - 4.4 Internet User

Figure 1: User profile typology in MARQUIS

Although the above typology and the possibility of its individualization certainly stand for the most advanced realization of user modelling in the framework of environmental information services, there is still some room for further work in this area. Thus, the typology does not reflect such important user groups as people with weak heart conditions and people doing outdoor sports. Members of both groups might need more targeted and distinct information than members of the groups reflected in the typology.

Recent studies in environmental medicine show that the influence of air pollution on the population is highly individual and can hardly be captured by generalized in terms of profiles. To account for this finding, a radically different approach to user modelling must be taken. In principle, for each user, a profile must be built up that reflects his/her individual information needs. The individualization will concern first of all the increased relevance of specific pollutant substances and the pollutant concentration thresholds beyond which a pollutant becomes relevant to the user in question. However, such individualization would require a tremendous effort not only on the environmental service side, but also on the medical side and is thus not realizable in the near future. Therefore, the goal in the field of user modelling for environmental information services must be a maximally detailed (but empirically justified) user profile typology. MARQUIS shows the way how this can be done.

4.2 Assessment and interpretation of raw data in MARQUIS

MARQUIS's assessment and interpretation module processes the raw data with respect to all topics listed in Section 2.2 above.

The interpretation procedure is a multistage procedure. In each stage, one of the above topics (i)-(iv) is addressed. The first step in the assessment procedure is the computation of the regional AQ indices. The indices are computed for each region. That is, the index for Finland is computed not only for Finland, but also for the other regions covered in MARQUIS; the same applies to the indices of the other regions. This supports the cross-border view on air quality in a given region. The indices are computed hourly as well as for the whole day. This allows the service to provide air quality information of a specific time of the day as well as for the day as a time interval.

The health relevance of a given pollution is determined by matching the measured concentrations or the index values with the concentration–health effect tables compiled, again, for each region.

As not every measured index value or pollutant concentration is important or interesting for the users, in the second step of the assessment procedure, very important points (VIPs) within the daily distribution curve are identified. Standard VIPs are daily and local maxima, respectively minima. In addition, the first

value of the day and the current value is a VIP. To determine the VIPs, a standard mathematical curve discussion is carried out. In order to avoid the identification of small (and thus irrelevant) local maxima/minima, certain smoothing strategies are applied.

National and European thresholds for the different pollutants are considered. To allow different region specific reports, the exceedance of these thresholds are provided in two ways. First, the threshold type and the start and end time of the exceedance are given, and, second, the degree of exceedance is provided in absolute and relative numbers.

An important feature of the MARQUIS-service is the availability of plausible explanations for measured and forecasted air pollutant concentrations. The correlation between air quality and meteorological conditions is the most important source for such explanations. A comprehensive range of potential correlations has been studied. The most relevant meteorological conditions include wind speed and direction, temperature, global radiation, precipitation, pressure difference between reference locations, relative humidity. External conditions such as day of the week (i.e., traffic), number of days since the last significant precipitation, etc. are also considered.

The work on content derivation from raw data within the MARQUIS-service constitutes pioneering work in this area. Its limitations lie rather in the user typology according to which the assessment is done and in the implementation, which is less modular than desired, than in the theoretical conception.

4.3 Generation of information in MARQUIS

In MARQUIS, a full-fledge text generator is applied. As pointed out above, full-fledged generation consists of document planning and generation proper.

MARQUIS's document planning module (Bouayad-Agha/Wanner this volume) involves, roughly speaking, two submodules: the submodule for content selection and the submodule for the determination of the discourse structure of the textual information to be generated. The content selection task consists mainly in the retrieval of the content relevant to the user in question according to his/her (individualized) profile from the content structure produced by the assessment and interpretation module on the one hand, and the choice of the information presentation mode (text, table, or graphic) for identified content chunks on the other hand. The discourse structure determination task consists in the derivation of the *rhetorical relations* between content elements and construction of a connected *discourse graph* from the content elements-rhetorical relation triples (which form elementary branches of the graph). For the repertoire and interpretation of the nature of rhetorical relations, MARQUIS draws upon the Rhetorical Structure Theory (Mann/Thompson, 1988).

MARQUIS's multimodal generation module consists of three submodules: a table generation submodule, a graphics generation submodule and text generation submodule – with the latter being the most demanding one (Bohnet this volume). For its realization, the MATE-generator is used (Bohnet, 2006). MATE is a graph transducer based generator that is based on the multi-stratal Meaning-Text linguistic model (Mel'čuk, 1988). Six strata are relevant in the context of MARQUIS: conceptual, semantic, deep-syntactic, surface-syntactic, topological, and surface-morphological strata. Generation consists in mapping an input conceptual structure retrieved from the text plan provided by the document planning module to the surface-morphological structure (= text) via the structures of the intermediate strata listed above. The mapping of the structures between two adjacent strata is realized by a distinct language-specific grammar. In order to generate information in a specific language, thus five grammars and the corresponding lexica must be developed. To facilitate the work of non-linguists, MATE offers a grammar development environment with a debugger, grammar inspector, interactive editors, etc. (Bohnet et al., 2000). Within MARQUIS, grammars for eight languages have been developed: Catalan, English, French, Finnish, German, Polish, Portuguese, and Spanish. For the organization of the grammatical resources is described, see (Lareau/Wanner, 2007).

A topic not targeted in depth in MARQUIS in connection with information generation is layout planning: how to place the information generated by the different modi adequately in order to guide the reader through the document.

5. Conclusions

Advanced environmental information services must contain modules for interpretation of the data, i.e., their conversion into content, document planning and information generation. The MARQUIS-service is certainly one of the first services that account for these requirements – despite some limitations that need to be solved in order to make it a truly next generation information service.

Bibliography

Bernd, Bohnet. (2006): Textgenerierung durch Transduktion linguistischer Strukturen. Köln: DISKI 298.

Bohnet, B. (et al.) (2001): AutoText-UIS – Automatische Produktion von Ozonkurzberichten im Umweltinformationssystem Baden-Württemberg, in K. Tochtermann and W.-F. Riekert (eds.); Neue Methoden für das Wissensmanagement im Umweltschutz, pp. 21-32.

Bohnet, B., Lareau, F., Wanner, L. (this volume): Automatic production of multilingual environmental information.

Bohnet, B., Langjahr, A., Wanner, L. (2000): Development Environment for an MTT-Based Sentence Generator. In Proceedings of the First International Conference on Natural Language Generation. Israel, Mitzpe Ramon.

Bouayad-Agha B., Wanner, L. (this volume): Text planning of air quality Information.

Lareau, F., Wanner, L. (2007): Towards a generic multilingual dependency grammar for text generation. In Proceedings of the Symposium "Grammar Engineering across Frameworks". CSLI: Stanford.

Mann, W.C, Thompson, S.A. (1988): Rhetorical Structure Theory: A theory of text organization. In: L. Polanyi (ed.) The Structure of Discourse, Ablex.

Mel'čuk, I. (1988): Dependency Syntax. Albany, NY: SUNY Press.

Nicklaß, D. (this volume): Addressee-tailored interpretation of air quality data.

O'Donnell, M., Mellish, C., Oberlander, J., Knott, A. (2001): ILEX: An architecture for a dynamic hypertext generation system. Natural Language Engineering 7:225-250.

Paris, C. (1993): User modelling in text generation. London: Frances Pinter.

Wanner, L. (et al.) (2007): From Measurement Data to Environmental Information - MARQUIS - A Multimodal AiR QUality Information Service for the General Public. In: Swayne A., Hrebicek J. (eds.): Environmental Software Systems – Dimensions of Environmental Informatics. Vol. 7.

Zukerman, I., Litman, D. (2001): Natural language processing and user modelling: Synergies and limitations. User Modeling and User-Adapted Interaction, 11:129-158.