Latent User Models for Online River Information Tailoring

Xiwu Han[1], Somayajulu Sripada[1], Kit (CJA) Macleod[2], and Antonio A. R. Ioris[3]
Department of Computing Sciences, University of Aberdeen, UK[1]
James Hutton Institute, Aberdeen; University of Exeter, Exeter, UK[2]
School of GeoSciences, University of Edinburgh, UK[3]
{xiwuhan,yaji.sripada}@abdn.ac.uk
kit.macleod@hutton.ac.uk
a.ioris@ed.ac.uk

Abstract

This paper explores Natural Language Generation techniques for online river information tailoring. To solve the problem of unknown users, we propose 'latent models', which relate typical visitors to river web pages, river data types, and river related activities. A hierarchy is used to integrate domain knowledge and latent user knowledge, and serves as the search space for content selection, which triggers user-oriented selection rules when they visit a page. Initial feedback received from user groups indicates that the latent models deserve further research efforts.

1 Introduction

Within recent decades, access to online river information has increased exponentially thanks to great progresses in data collection and storage technologies employed by hydrological organizations worldwide (Dixon, 2010). Local residents nearby rivers and those engaged in river related activities are now much better informed and more engaged with data providers than decades ago. However, organizations such as SEPA (Scottish Environment Protection Agency), CEH (Centre for Ecology and Hydrology), EA (Environment Agency) in UK, and quite a few Canadian and Australian ones are working to improve the presentation of river information further. Many of these data providers, who are mostly government agencies, provide descriptive texts along with archived data of flow, level, flood and temperature along with their graphs and/or tables. A typical example of linguistic description from the EA website is shown below:

> The river level at Morwick is 0.65 metres. This measurement was recorded at 08:45 on 23/01/2013. The typical river level range for this location is between 0.27 metres and 2.60 metres. The highest river level recorded at this location is 6.32 metres and the river level reached 6.32 metres on 07/09/2008.[1]

The above descriptive text could vary to some extent according to different river users. For instance, it may provide information perceived as good news by farmers whilst other users e.g. canoeists or paddlers may interpret the information as bad news for their activity. Such tailored information provision promotes communication efficiency between stakeholders and the relevant government offices (Macleod et al., 2012). We explored data-to-text techniques (Reiter, 2007) in promoting online river information provision. Our engagement activities with river stakeholders showed that there could be great difficulties in specifying user groups for online river information tailoring. First, the relations between domain knowledge and user knowledge are difficult to be acquired due to domain sensitive challenges. Second, for online communication, the issue that users themselves sometimes are not sure about their tasks further hinders user modeling. This paper proposes an alternative approach of latent user models, instead of directly asking users to indicate what they are interested in.

2 User Modeling Problem

It has long been argued in NLG research that contents of generated texts should be oriented to users' tasks and existing knowledge. User models are usually employed for the tailoring task. However, user models may not be easily acquired. Reiter et al (2003a) claimed that no NLG system actually used detailed user models with non-trivial numbers of users. Most commercial

[1] http://www.environment-agency.gov.uk/homeandleisure/floods/riverlevels/120694.aspx?stationId=8143

NLG systems would rather do with very limited user models, and examples are STOP (Reiter et al., 2003b), SUMTIME-MOUSAM (Sripada et al., 2002), and GIRL (Williams, 2002).

Recent research on user modeling falls into roughly three categories, i.e. explicit, implicit and hybrid approaches[2]. All approaches start with knowledge acquisition. Explicit models then define a finite number of user groups, and finally generate tailored texts for users to choose from, or choose to generate for a unique group at each time, e.g. (Molina, 2011 and 2012). Implicit models, e.g. (Mairesse and Walker, 2011), then construct a framework of human computer interaction to learn about the values of a finite set of features, and finally generate tailored texts according to the intersection between domain knowledge and feature values. Hybrid models, e.g. (Bouayad-Agha et al, 2012) and (Dannels et al, 2012), specify both a finite set of user groups and a human computer interaction framework, and finally classify online users into defined groups for tailored generation.

3 Latent User Models

Online river information tailoring involves a website, such as SEPA's, which provides map based (or text based) searchable river information [3]. The NLG task is to generate user-oriented texts while users are navigating the website. Both explicit and implicit user models can be employed for online river information tailoring. A finite set of user groups could be defined according to river-related activities, such as flooding, fishing, canoeing, etc. along with a set of features such as level trends, temperature ranges, etc. Then an interactive navigation mechanism could ask a user to either choose a group or tailor his/her own parameters, and relevant texts can be generated thereafter.

Unfortunately, our engagement activities with stakeholders showed that it is almost impossible to define user models using mappings from river-related activities to river data features. Furthermore, frequent users are reluctant to spend time on specifying their preferences before viewing the river information. For such an NLG task, the uncertainty comes not only from a large variety of river users and stakeholders, but also from the issue that users themselves sometimes are not sure of what data features are associated with making decisions about their activities.

Our efforts on dealing with NLG domain knowledge and user models brought about the idea of extending domain knowledge to statistically cover user knowledge, without explicitly defining user groups or implicitly modeling potential users. We argue that non-trivial number of uncertain users can be dynamically and statistically modeled by integrating a module for web mining and Google analytics into the NLG pipeline system. We regard these statistically established models as latent since they are hidden beneath the domain knowledge, and the latent variable of typical users is linked to river data types and river related activities.

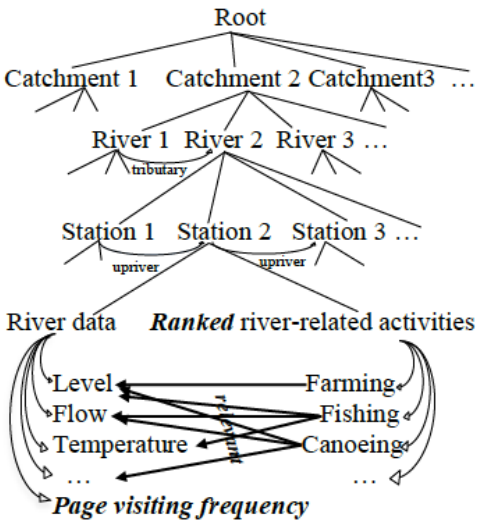

Figure 1. Domain Knowledge with Latent Models

The domain knowledge and latent user models are constructed as a whole in a hierarchical structure, as in Figure 1. We technically maintain this hierarchy as an ontology based on existing approaches e.g. (Bontcheva, 2005; Bouayad-Agha et al, 2012). The general part of the main frame was extracted from hydrology or environment websites, such as SEPA, CEH and EA, with the view that these websites were deliberately established hierarchically by manual work of domain experts in the fields of hydrology, ecology and/or geology. This part serves as the center of our domain knowledge, which starts with a root node and branches to river catchments, rivers, river stations and river data, while river data consists of water level, water flow, water temperature, etc. There are also some non-hierarchical relations embedded, namely the tributary relation between rivers, the upriver relation between river stations, and the relationship between certain river data and river related activities. In addition

[2] Note the difference between NLG and HCI user models. The former tailor the output of NLG systems, while the later tailor the systems themselves.

[3] http://sepa.org.uk/water/river_levels/river_level_data.aspx

to the time series on the status of the rivers, other information is integrated offline. Then, the domain knowledge was extended to cover potential users' knowledge and online visiting behaviors. The extended information, or the latent user models, as denoted in italic fonts in Figure 1, includes three parts, i.e. the webpage visiting frequency, the relevance degrees between certain river data and river related activities, and the ranking of popularities of river-related activities for each river station.

Our extension process includes three stages, i.e. web mining, Google analytics, and engagement activities. At first, basic and rough information about river stations was statistically gathered by using free or trial version web mining tools, such as spiders and crawlers, and corpus analysis tools. For all combinations of elements respectively from each pair of columns in Table 1, we simply count the tokens of co-occurrence within an empirical window of 10 words. For the co-occurring tokens between a given river station and related activities, the top five tokens were selected by filtering according to one threshold on co-occurrence frequencies and another threshold on frequency differences between adjacent ranked types. For the co-occurring tokens between a given activity and river data type, relevant tokens were chosen by only one threshold on the co-occurrence frequencies. Finally, the co-occurring types of river stations and river data with high frequencies were used to fine-tune the previously acquired results, supposing that some river stations seldom or never provide some types of river data.

River Stations	Related Activities	River Data Type
Aberlour	Farming	Level
Aberuchill	Fishing	Flow
Aberuthven	Canoeing	Temperature
Abington	Swimming	Width
Alford	Kayaking	Rainfall
Allnabad	Rowing	Wind
Almondell	Boating	Pollution
Alness	Research	Birds
Ancrum	Education	Animals
Anie	Hiking	Fishes
Apigill	Cycling	…
Arbroath	…	…
…	…	

Table 1. Basic Domain Knowledge for Extension

We further had the statistically acquired results complemented and modified by Google analytics data for river websites and engagement activities with domain experts and users. Google analytics provided us with webpage visiting frequencies for each hydrological station, and contributed to the ranking of river-related activity for a given station. Knowledge gathered from engagement activities, such as semi-structured interviews and focus groups, was mainly used to confirm the statistically gathered information during the first two stages (as well as refine our overall understanding of data demands, water-related activities and perception of existing communication tools). For example, flood warning information was moved up in the ranks since over 5 million people in England and Wales live and work in properties that are at risk of flooding from rivers or the sea[4] (Marsh and Hannaford, 2007). Our present research is limited to rivers in Scotland, involving 107 river catchments, 233 rivers, and 339 river stations. The webpage visiting frequencies for these stations were gathered from Google analytics data for the website of SEPA[5]. The page visiting frequency for each river station is represented by a time series with yearly periodicity, and each period includes 12 numeric elements calculated by dividing the number of monthly visiting times of the station by the total number of monthly visiting times of all river stations.

4 NLG for Online Tailoring

Our NLG pipeline system takes numeric data of a given river station as input, and outputs a tailored description for that river station. The system analyzes data of water level, flow, and temperature as similar to time series analysis tasks presented in (Turner et al., 2006). Then, the analyzed patterns are interpreted into symbolic conceptual representations, including vague expressions, which might facilitate users' understanding (van Deemter, 2010). SEPA defines normal ranges for river levels and we use these definitions in our computations to generate vague expressions. For content selection, we define five sets: $\mathbf{S} = \{s_1, s_2, \ldots\}$ the set of stations; $\mathbf{A} = \{a_1, a_2, \ldots\}$ the set of activities for a given station; $\mathbf{D} = \{d_1, d_2, \ldots\} = \{\{d_{11}, d_{12}, \ldots\}, \{d_{21}, d_{22}, \ldots\}, \ldots\}$ the set of river data sets for a given station; $\mathbf{AD} = \{a_1d_1, a_1d_2, \ldots, a_2d_1, \ldots\}$ where a_id_j refers to information from the interpretation of an activity a_i under the condition of data d_j; and **SAD** an overview on one station. For a river station, using the domain knowledge hierarchy, which embeds la-

[4] http://www.environment-agency.gov.uk/homeandleisure/floods/default.aspx.
[5] http://www.sepa.org.uk.

tent user models implicitly (Figure 1), we select **A** ∪ **D** ∪ **AD** ∪ **SAD** as the initial contents.

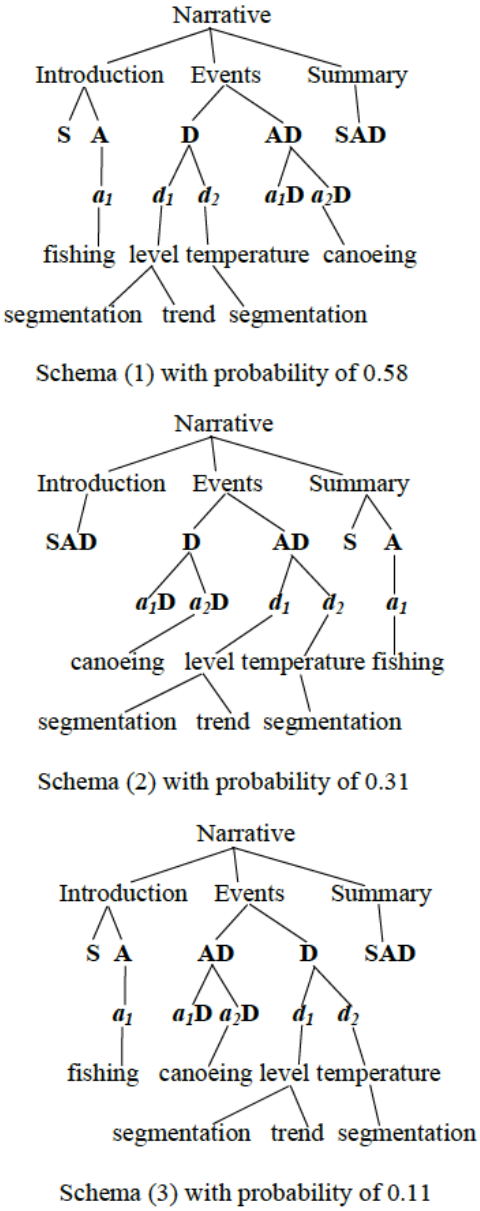

Figure 2. Statistical Schemas

A schema-based approach was employed for document planning. Each schema at the high level is made up of three components: Introduction, Events and Summary. Each of these components has its own substructure as shown in examples in Figure 4. With the estimated probabilistic distribution we generate schemas for a station based on its popular activities. We then tailor the text by randomly selecting from users' favorite vocabulary, which was acquired from online corpus for different river-related activities. Other words for structural purposes are dependent on certain schemas. Realization was performed using the simpleNLG library (Gatt and Reiter, 2009), and some generated examples are listed in Table 2.

Schema (1)	The ***Tyne at Nungate*** boasts its excellent ***salmon catches***. Now with ***medium steady*** water level and comparatively ***low water temperature***, many people want to ***fish*** some ***salmons*** in pools between the rapids or experience whitewater ***rafting*** within them, which makes the periphery of ***Nungate a hot spot***.
Schema (2)	The periphery of ***Tyne at Nungate*** poses a hot spot now, where many people are ***fishing*** or ***canoeing*** while appreciating the ***medium steady*** water level and comparatively ***low water temperature***. No wonder ***Nungate*** can boast one of the best ***salmon catching*** places.
Schema (3)	The ***Tyne at Nungate*** boasts its excellent ***salmon*** catches. Many people may now ***fish*** or ***canoe*** there thanks to the ***medium steady*** water level and comparatively ***low water temperature***, making the periphery of ***Nungate a hot spot***.

Table 2. Some Tailored NLG Examples (Italic fonts denote the tailored lexical realization)

5 Initial Feedback and Conclusion

This research is still underway and a thorough evaluation is still pending. We have received valuable feedback from small user groups. Supportive examples are: a. An overview about popular river stations can help users' further exploration of information to a significant extent; b. A general comprehension for a given river station can be more easily built up by simply reading the generated descriptions, than by solely reading the data and its related graphics; c. Along with the graphics, the generated descriptions can improve the communication efficiency by a large degree. Examples recommending further improvement/focus include: a. Schemas filled in with acquired vocabulary sometimes endow the generated document a syntactically and/or semantically unexpected flavor; b. Established users demand more linguistic varieties than new users.

Present feedback implicates that latent user models deserve further research. Our future efforts will focus on a. extending the domain knowledge to cover all river stations, b. developing generic methodology for acquiring latent user models for other online NLG tasks (e.g. generating descriptions of Census data), and c. integrating an automatic update of latent models.

Acknowledgement

This research is supported by an award from the RCUK DE programme: EP/G066051/1. The authors are also grateful to Dr. Rene van der Wal, Dr. Koen Arts, and the three anonymous reviewers for improving the quality of this paper.

References

K. Bontcheva. 2005. Generating Tailored Textual Summaries from Ontologies. *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, Vol. 3532, pages 531-545. Springer-Verlag.

N. Bouayad-Agha, G. Casamayor, Simon Mille, et al. 2012. From Ontology to NL: Generation of Multilingual User-Oriented Environmental Reports. *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science Vol. 7337, pages 216-221. Springer-Verlag.

Dana Dannells, Mariana Damova, Ramona Enache and Milen Chechev. 2012. Multilingual Online Generation from Semantic Web Ontologies. *WWW 2012 – European Projects Track*, pages 239-242.

H. Dixon. 2010. Managing national hydrometric data: from data to information. *Global Change: Facing Risks and Threats to Water Resources*. Wallingford, UK, IAHS Press, pages 451-458.

A. Gatt and Ehud Reiter. 2009. Simplenlg: A Realization Engine for Practical Applications. *Proceedings ENLG-2009*, pages 90-93.

K. Macleod, S. Sripada, A. Ioris, K. Arts and R. Van der Wal. 2012. Communicating River Level Data and Information to Stakeholders with Different Interests: the Participative Development of an Interactive Online Service. *International Environmental Modeling and Software Society (iEMSs): International Congress on Environmental Modeling and Software Managing Resources of a Limited Planet*, Sixth Biennial Meeting, Leipzig, Germany. R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.) pages 33-40.

Francois Mairesse and Marilyn A. Walker. 2011. Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits. *Computational Linguistics*, Volume 37 Issue 3, September 2011, pages 455-488.

T. J. Marsh and J. Hannaford. 2007. *The summer 2007 floods in England and Wales – a hydrological appraisal*. Centre for Ecology & Hydrology, UK.

M. Molina. 2012. Simulating Data Journalism to Communicate Hydrological Information from Sensor Networks. *Proceedings of IBERAMIA*, pages 722-731.

M. Molina, A. Stent, and E. Parodi. 2011. Generating Automated News to Explain the Meaning of Sensor Data. In: *Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS*, vol. 7014, pages 282-293. Springer, Heidelberg.

Ehud Reiter, Somayajulu Sripada, and Sandra Williams. 2003a. Acquiring and Using Limited User Models in NLG. In *Proceedings of the 9th European Workshop on Natural Language Generation,* pages 13-14, Budapest, Hungary.

Ehud Reiter, Roma Robertson, and Liesl Osman. 2003b. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2), pages 41-58.

Ehud Reiter. 2007. An Architecture for Data-to-Text Systems. *Proceedings of ENLG-2007*, pages 97-104.

S. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2002. Segmenting time series for weather forecasting. *Applications and Innovations in Intelligent Systems X*, pages 105-118. Springer-Verlag.

R. Turner, S. Sripada, E. Reiter and I. Davy. 2006. Generating Spatio-Temporal Descriptions in Pollen Forecasts. Proceedings of EACL06 poster session, pages 163-166.

K. van Deemter. 2010. Vagueness Facilitates Search. *Proceedings of the 2009 Amsterdam Colloquium*, Springer Lecture Notes in Computer Science (LNCS). FoLLI LNAI 6042.

Sandra Williams. 2002. Natural language generation of discourse connectives for different reading levels. In *Proceedings of the 5th Annual CLUK Research Colloquium*, Leeds, UK.