Web as Corpus Supporting Natural Language Generation for Online River Information Communication

Xiwu Han
Department of Computing Science
University of Aberdeen
Aberdeen, UK
+44(0)1224275441
xiwuhan@abdn.ac.uk

Antonio A. R. Ioris
School of GeoSciences
University of Edinburgh
Edinburgh, UK
+44(0)1316519090
a.ioris@ed.ac.uk

Chenghua Lin
Department of Computing Science
University of Aberdeen
Aberdeen, UK
+44(0)1224272306
chenghua.lin@abdn.ac.uk

ABSTRACT
Web as corpus for NLP has been popular, and we now employed web as corpus for NLG, and made the online communication of tailored river information more effective and efficient. Evaluation and analysis shows that our generated texts were comparable to those written by domain experts and experienced users.

Categories and Subject Descriptors
H.3.5 [**Online Information Services**]: Web-based Services – *Miscellaneous*

General Terms
Performance, Experimentation, Human Factors

Keywords
Web; Corpus; NLP; NLG; river information communication

1. INTRODUCTION
Motivated by web as corpus for NLP [3], this research investigated applying web as corpus for Natural Language Generation (NLG) to effectively communicate online river information to users. NLG is a knowledge-based task [7, 8], requiring knowledge of language, domain and users. Traditionally, experts, users and application specific corpus have been the main sources for acquiring the required knowledge, but this can be very expensive and time consuming. The current work evaluated the hypothesis that the web as NLG corpus can be a good knowledge source.

2. Related Work
Recent NLG research on user knowledge acquiring falls into roughly four categories, i.e. explicit, implicit, hybrid and latent approaches [2]. All approaches start with knowledge acquisition. Explicit models will 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. Implicit models will then construct a framework of human computer interaction to value a finite set of parameters for user information, and finally generate tailored texts according to the overlapping between domain knowledge and user information. Hybrid models will 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. Latent models provide an alternative solution to the problem of new unknown users.

Online river information communication has also been developed in automatically generated journal news [5, 6], which is intended for different users, including residents in certain catchment area, hydrological experts, and relevant government offices. Three explicit user groups were defined in terms of communication goals, i.e. flood risk management, water management, and sensor validation. All three kinds of news are generated for each geographic area, and different users in this case can choose to read about their own concerned blocks in the electronic journal. The news is generated offline before users surfing river webpages, and not tailored to the special interest of a particular user group concerning a particular river station. In strict sense, the generated journal news does not provide pure online communication. Our research, in contrast, can generate narrative descriptions online while users surfing river webpages of their concerned river stations, and the information can be tailored both to typical users and a particular river or a hydrological station.

WWW'15 Companion, May 18-22, 2015, Florence, Italy.

http://dx.doi.org/10.1145/2740908.2745938

3. Mining NLG Templates
We started from a small seeding set of Scottish river names, names of gauging stations, and eight common keywords. This paper mainly focuses on the river of Dee (Grampian) and its gauging stations. We used each seeding keyword as a query while searching within English webpages of UK Broadsheet Newspapers, UK Tabloid Newspapers, BBC News, and Wikipedia. Frequent collocations were then extracted and sorted after excluding stop words. With the extended keywords as queries, we collected river related corpus of 1,956 documents. The Stanford NLP toolkit of CoreNLP 3.4 [4] was employed to process the gathered corpus for syntactic parse trees, dependency trees, and named entity recognition results. Two kinds of keyword collocations were acquired from dependency trees of the corpus. They are collocations of keyword pairs, and collocations of word pairs between keywords and other words. Both river activities and river information are regarded as attributes for defining a river or a gauging station, or as topics in river descriptions. Besides, river information also describes the relevant situations for certain river activities. By mapping rivers and river stations with activities, we found out what river users might most likely engage in along a river or nearby a station. By mapping river activities and river information, we found out how differently river activities were described in our gathered corpus.

We used templates to model river users and generate tailored narrations about river situations. In NLG research, templates are often employed as linguistic surface structures, which may contain gaps, and become well-formed output results when the gaps are filled with linguistic structures without gaps [1]. For all keyword dependency collocations, we extracted sub-trees in the corresponding syntactic parse trees by mapping rules from dependency types to phrase structures. Finally, each segment was generalized by labeling keywords with keyword types. To build well-formed linguistic surfaces, we generated syntactic templates by first integrating relevant template segments into each of the

discourse structure candidates, then filtering the candidates by a simple grammar, and finally ranking the rest by their probabilities. The simple grammar only includes two product rules: "S => NP VP | NP VP NP", and "?/NP => ? NP". Here, "?/NP" refers to any terminal or nonterminal in a template segment that needs a NP phrase to be complete. For example, the segment "(S *rainfall* increased/NP)" needs some NP phrase like "the river level" to be complete. Each discourse structure may be realized with a few syntactic templates, such that all relevant information types are covered. For ranking, we used the joint probability of all template segments involved in the realized structure. These syntactic templates were employed to model river users and tailor the linguistic surface for information communication.

4. Natural Language Generation (NLG)

There are four main parts in the prototype, i.e. user models, HCI, background knowledge, and NLG. Our present user models cover five groups: new user, fishing, canoeing, flooding and others, which were learned from public web corpus and can be easily modified or updated by mining more corpus. The HCI part provides a simple mechanism either for a user to choose a model or for the system to assign a user with the most likely model based on his/her visiting history. Our present system simply assigns a new unregistered visitor with the new user model, and a returning visitor with his/her previous model. Background knowledge includes hydrological and geographical knowledge about Scottish rivers and stations, and the nearby popular river related activities. The NLG part takes numeric input data, analyzes data for messages [8] and outputs short English texts tailored to different user groups. Both user models and background knowledge are employed to select content for NLG, namely filtering the input data at first and then selecting proper messages. The templates are already proper discourses for river information communication. To ensure some linguistic variety, templates are chosen randomly with regard to their probabilistic distribution within a user model.

5. Evaluation

Domain experts and river users were engaged in our evaluation. The generated texts were evaluated and analyzed against descriptions written by domain experts and experienced users. For users to rate the texts, three measurements were employed: a. "helpful" (*He*) – how the text could help the user's decision making in relation to his/her river activity, ranging from "irrelevant" to "very helpful"; b. "clear" (*Cl*) – whether the information is readily understandable, ranging from "not at all clear" to "very clear"; and c. "concise" (*Co*) – whether the text is brief and to the point, ranging from "far too wordy" to "concise". Each measurement was assigned with a scale from one to nine. We invited 42 users for the text rating evaluation. Each user was asked to rate 36 texts for the three measurements and optionally comment according to his/her gut reaction. Statistics about the rating points are listed in Table 1. We find that to most users our generated texts for the Others user group were received as more helpful and clearer than descriptions written by domain experts or experienced users. One way ANOVA F-test with $F_{crit}(6, 252) = 2.22$ at $\alpha = 0.05$ showed that the differences were statistically significant, for "helpful" (*He*) $F = 4.06 > 2.22$, P-value = 0.001, and for "clear" (*Cl*) $F = 11.11 > 2.22$, and P-value = 0.0002. However, on the measurement of "concise" (*Co*) two human writers outperformed the NLG system significantly, with $F_{crit}(6, 252) = 2.22$ at $\alpha = 0.05$, $F = 78.29 > 2.22$, and P-value = 0.0004. Our mined models also lead to generally smaller standard deviations for most rating distributions than human writers. Therefore, our system based on web corpus is comparable to human writers of domain experts and experienced users, and performs in a more stable way.

Table 1. Some statistics about the rating points

		Baseline	New	Others	H1	H2	H3
He	Median	6	6	7*	6	6	6
	Mean	5.84	5.85	6.4*	5.83	6.09	5.72
	STDEV	1.86	2.15	1.65*	1.98	1.86	2.21
Cl	Median	7*	6	7*	6	6	6
	Mean	6.66*	6.08	6.64	6.13	6.23	5.68
	STDEV	1.67	1.74	1.55*	1.95	1.83	1.90
Co	Median	6	6	6	7*	7*	4
	Mean	5.89	5.35	5.96	6.74*	6.67	3.93
	STDEV	1.91	1.80	1.78	1.67*	1.81	2.14

6. Conclusion

In this paper, web corpus mining played a central role in acquiring user relevant knowledge and supporting an NLG-based prototype system. Evaluation and analysis shows that our generated texts were comparable to those written by domain experts and experienced users, and the performance of our prototype system is more stable and more economic than human writers.

7. ACKNOWLEDGMENTS

This research is supported by an award from the RCUK DE programme: EP/G066051/1.

8. REFERENCES

[1] Deemter, K. V., Krahmer, E. and Theune, M. 2005. Real vs. template-based NLG: a false opposition? *Computational Linguistics*, 31, 1(March 2005), 15-24.

[2] Han, X., Sripada, S., Macleod, C. and Ioris, A. 2014. Latent User Models for Online River Information Tailoring. *The 8th International Natural Language Generation Conference*. 133-137, Philadelphia, Pennsylvania, 19-21 June 2014.

[3] Liu, Vinci, and James R. Curran. 2006. Web Text Corpus for Natural Language Processing. *Proceedings of EACL 2006*, 233-240.

[4] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.

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

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

[7] Reiter, E. and Dale, R. 2000. *Building natural language generation systems*. Cambridge University Press, Cambridge, UK.

[8] Reiter, E. 2007. An Architecture for Data-to-Text Systems. *Proceedings of ENLG-2007*, 97-104.