# MITACS Seed Project Proposal

# Multi-format environmental information dissemination

Guy Lapalme
Philippe Langlais
*RALI-DIRO*

Pascal Vincent
*LISA-DIRO*

Université de Montréal

April 24th 2009

# DESCRIPTION OF PROJECT

This project will explore new ways of customizing and translating the mass of daily information produced by Environment Canada (EC). This information in digital format is later transformed into weather and environmental forecasts, warnings and alerts that must be broadcast in real-time in at least two languages, in many different formats and in a way that takes location into account.

Some of the available information cannot always be made available in a timely or accessible manner to Canadians because of the need for manual interventions to summarize it and translate it. Given the importance of weather and environmental information in so many spheres of activity, this project explores innovative ways for combining, condensing and displaying the specific types of data encountered in the environmental domain. It combines data-mining techniques in order to extract important information, with cognitive and linguistic methods for displaying it to the users in a meaningful way.

The research challenges of this project are the following:

- Selective and customized environmental information display
- Data-mining of relevant information patterns;
- Machine translation of weather forecasts.

This project addresses a real-life problem involving multi-format content generation, which arises from the practical needs of automatic weather and environmental information from meteorological databases. In the general case, these problems are very difficult to solve, but given that the discourse domain of weather information is relatively limited, we expect that they will remain computationally tractable in the two-year time frame of this research project. The close collaboration with our research partners at Environment Canada should enable us to focus on the most important *low hanging fruit*. Should the results of this seed project be successful, we hope to establish a longer-term cooperation with EC.

**Context of the work**

Figure 1 shows, from left to right, a *simplified view* of the information flow at EC, going from the numerical weather prediction (NWP) to the French and English weather bulletins issued regularly, as well as the warnings that are produced when exceptional weather or environmental special events occur, e.g. storms, hail, freezing rain, etc. These inputs and outputs are shown in bold in Figure 1.

The NWP model produces a lot of data every 12 hours: 6 Mb for each of the 5 regions of Canada and 9 Mb for global information for all Canada. This data is stored in different formats, one of them being *Scribe matrices* retaining estimated weather parameters such as wind speeds, pressures, temperatures, precipitations and others at 800 points across Canada. These matrices are processed by Scribe, a system that allows meteorologists to combine and correct information coming from the numerical model in order to provide a gamut of aggregated information. The most visible of these are the French and English weather reports that are generated automatically by a module of the Scribe system three times a day. Meteorologists can also issue weather warnings, most often in English, that are translated by a team of human translators aided by a machine translation system. Another output of Scribe is a set of XML files in a format called Meteocodes that give detailed information about weather events in each region or groups of geographical points. Currently, this data is made available to other clients of EC (e.g. The Weather Channel) that are interested in developing their own applications or forecasts.

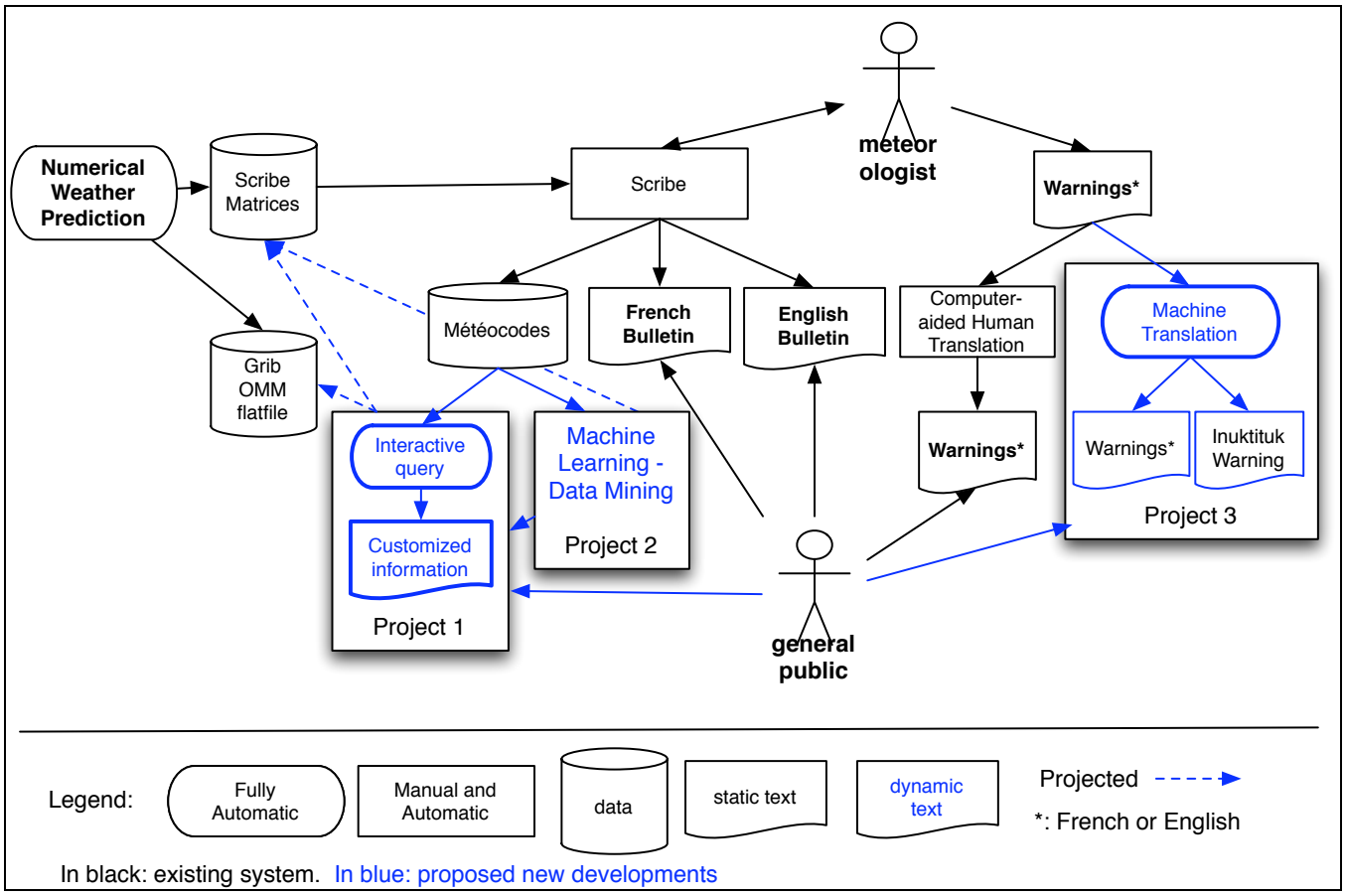


Figure 1: Simplified information flow at Environment Canada.
The modules and outputs to be developed in this proposal are shown in blue in the three Project boxes.

## Projects to be developed in this proposal

### 1: Selective and customized environmental information display

As previously mentioned, EC produces vast amounts of information: 26Mb of Meteocodes in XML every 12 hours. Selective and customized information display would allow EC to provide the Canadian public with much more focused forecasts, in both time and space, than the ones it currently produces, which are limited to the few dozens words found in the regional weather bulletins (about 1000 bulletins a day).

EC has already developed a text generator to produce the daily information on each region of Canada in both French and English. But this information is not as detailed as it could be, because the aggregation process requires limiting the number of different bulletins. For example, this aggregation results in precipitation probabilities rather than precise timing for the rain, and anticipated ranges in temperature that are wider than necessary. From the information in the Meteocodes, we want to develop a weather bulletin generator for a given address or postal code. It would be impractical to generate all these bulletins in advance (in both French and English); the great majority of them would never be read anyway. Further more, the regional weather information must also be made available in different output modes: graphical, Web, on weather radio and automated answering machines. An important goal of our project is to study the development of innovative approaches for conveying relevant meteorological information based on geographical and time-dependent aggregation.

As the Meteocodes are already in XML format validated by an XML Schema, we are sure that the input is easily parsable and we will focus on determining the most appropriate way of presenting the data in a meaningful way depending on the type of output device.

As we will be dealing with XML data, we will first study the use of XML StyleSheets (XSLT) to select and produce Web pages, but given the sheer size of the data, we do not expect that a simple selection and display will be enough. We will have to devise special purpose techniques for aggregating data both spatially and within time. In order to keep the project to a manageable size, EC has selected a group of 23 Canadian cities (out of the 800 alluded above) that account for about 75% of the population and for which we will develop customized information modules.

We intend to build on our previous experience in the integration of text and graphics in the context of geographical information. In (Fasciano & Lapalme, 2000), we showed how to achieve an integrated generation of text and graphics in statistical reports by simultaneously considering the writer's goals, the types and values of the variables to be presented, and the relations between these variables. We also incorporated good design rules that have a direct influence on the reader's perception of a report. This work has illustrated some of the complexities that arise even in the case of seemingly simple information graphics. We intend to apply the same principles to the domain of environmental information graphics in which the spatial component plays an important role.

Generating parallel text in many languages from one source is quite appealing from a conceptual point of view and has been cited as one of the potential applications for natural language generation (Reiter and Dale, 2000). Some practical systems have already been developed e.g. (Goldberg et al., 1994). These generation systems often require that humans select templates to organize the reports. Given the wide variety of devices available (Web, text, television, hand-held computers etc.), tailoring the Meteocode for each output device becomes prohibitive. On the other hand, the same information should not be presented in exactly the same way in text to be read on the radio, to be presented on television, on a cell-phone, on a Web site, on weather radio and on automated answering machines. Each type of device brings its own constraints and presents new opportunities. But care must be taken that the meaning of the information remains intact.

**2: Statistical Machine Learning and Data-Mining of relevant information patterns**

The raw data made available to us by Environment Canada contains a huge amount of detailed, fine grained (both spatially and in time) multi-dimensional information. This includes predictions of temperature, cloud coverage, precipitation, wind speed and direction, atmospheric pressure, humidity, air quality, etc... Initial predictions are produced by numerical simulation of a meteorological model that outputs predicted values for the relevant variables at each point of a fine grid across Canada, at different altitude levels in the atmosphere, for the next 36 hours. The prediction of this model, refreshed every 6 hours, is depicted as the *Numerical Weather Prediction* box on the above diagram. The result, exported as *Scribe Matrices*, can be thought of as a function $f^{\text{model}}(x,y,z,t)$ that associates a vector of predicted values for the variables of interest (temperature, pressure, etc...) at every point $u$ on a discrete fine grid, where $u=(x,y,z,t)$ is longitude, latitude, altitude, and time coordinates respectively. For a coarser grid (a subset of the points of the fine grid), these initial predictions are **examined and corrected** by hand by meteorologists familiar with each area. It is these corrected predictions that are output by the *Scribe* system as localized bulletins on the one hand and as Meteocodes, in XML format on the other. The Meteocodes can again be thought of as a function $f^{\text{meteocode}}(x,y,z,t)$ associating a vector of predicted

values of meteorological variables to a spatio-temporal coordinate, but this time on the coarser grid, due to limited human resources. It is these XML Meteocodes that are available to us.

A major goal of this project is to enable a user to access the wealth of information contained in these Meteocodes, and present it in a way that he can customize to his needs. This will require employing, researching and extending statistical machine learning and data-mining techniques. At a minimum, we want the user to be able to specify a geographical location of interest, as latitude and longitude or as a postal code, and be able to view the forecast evolution of all meteorological quantities of interest selected by him, possibly at different atmospheric levels, for the next 36 hours at that location. This can be presented to him, at his option, either as numerical tables, as graphical plots of time-series, or as an automatically generated descriptive text. A first difficulty, is that the high-quality, human-corrected $f^{\text{meteocode}}$ predictions available to us are only defined on a coarse grid, so we may need to interpolate or extrapolate the available predictions, in a smart way, to the requested specific location. This will yield a function *f*(u)* defined not only on a discrete grid, but on the whole territory. Gaussian Process Regression (Rasmussen and Williams 2006), also known in geostatistics as as *kriging* (Stein 1999, Cressie 1993), appear well suited for this task. Gaussian Processes allow to define a prior distribution over functions *f(u)*, that takes into account the covariance between the values at different locations. The covariance is given by a kernel function *k(u,v)* of the coordinates of any two locations *u* and *v*. A typical form for such a kernel is e.g.

$$k(u,v) = \gamma + \alpha \exp\left(-\frac{1}{2}\sum_i \eta_i (u_i - v_i)^2\right) + \beta \sum_i u_i v_i$$

Its (hyper)parameters $\alpha$, $\beta$, $\gamma$, and $\eta$ can be learnt based on historical data of the predictions made at the grid points. We will also investigate alternative forms of kernels with higher capacity. Combining the Gaussian prior on functions and the "observed" values given by $f^{meteocode}$ on coarse grid points, one obtains a posterior Gaussian distribution for the predicted values at any non-grid point of the space. In the future, Environment Canada also plans to make accessible the predictions of the model on the fine-grid ($f^{model}$). When these will become available, the problem of computing a good prediction on any non-grid point will present itself differently: what we then have is a fine grid of raw (uncorrected) model predictions, and a coarse grid of high quality predictions corrected by meteorologists. This makes for an interesting machine learning research problem. One approach would be to extend or adapt the standard Gaussian Process model to handle this specific setting. Another more original approach would be to learn, from historical data collection, a statistical model of the corrections made by human meteorologist across different raw prediction conditions and geographical areas. For this we may build on our expertise in artificial Neural Networks and in particular recent advances (Vincent et al. 2008) for successfully training so-called *deep networks* able to model more complex functions of high dimensional data. Whatever the successful approach retained, it will yield a function *f** that can output a quality prediction at any precise location that is off the coarse grid (e.g. above a farmer's field).

In addition to accessing precise time-series of meteorological variables at a single specific spatial location, it will also be possible to produce maps of the values of variables across a whole *user-specified region* at a given time. Whatever its format (graphical time series, graphical map or text description), a display of information can be made much more valuable by drawing the user's attention to less usual phenomena and conditions. For this we plan to investigate the use of simple local kernel density estimation techniques (Vincent and Bengio, 2003) for outlier detection. Outputting time series of environmental variables in the form of a text description is another challenging task. One starting point

will be the results of the SumTime project (Reiter et al. 2005, Yu et al. 2007), a system for generating English summaries of numerical time-series by integrating time-series analysis techniques in the context of natural language generation. More generally, we want to address automatic generation of a text description that summarizes specific meteorological conditions across a *region*. This is an unsolved problem. In our view, any such summarization will first require performing a spatio-temporal clustering based on similarity of conditions, before we can report the found clusters+conditions. We thus intend to adapt spectral clustering algorithms (Ng et al. 2000) and integrate them with natural language generation.

**3: Statistical Machine translation of weather forecasts**

Although many weather reports and forecasts from Environment Canada are routinely generated automatically (see above), there are still many cases, especially in severe weather conditions, in which meteorologists write their forecasts directly either in English or French. These reports must then be translated in the other official language before being broadcast. This human translation is costly, time consuming and must be carried out under severe time constraints, given the emergency situations often associated with this type of information.

Since the 1970's, Environment Canada has been a leader in the integration of symbolic machine translation in its weather bulletin broadcasting process. It still uses a variation of the original technology for translating weather alerts. But given the recent and impressive progress in statistical machine translation (SMT), EC wants to investigate how this new technology could be applied for these types of weather alerts and also to study how new languages (e.g. Inuktitut or Chinese) could be accommodated more rapidly than the current painstaking process of handcrafting new grammar rules.

We have already done some preliminary experiments in the translation of these types of weather bulletins. In (Langlais et al., 2005), we compared various ways of implementing corpus-based approaches for the translation of weather reports, for which huge amounts of bitexts are now available. We observed that a straightforward memory-based approach can already produce good results, owing to the highly repetitive nature of the weather forecast domain. We found that a phrase-based SMT engine is even better suited to translate previously unseen sentences, and registered further improvements after applying a rescoring layer. Finally, combining a translation memory and an SMT engine yielded significant overall improvements.

We also examined another extension of this technology to a more challenging task: the translation of weather alerts. Here, however, our approaches were unable to achieve the same level of success without further adaptation. Nevertheless, we did confirm that a combination of translation memory and a statistical phrase-based engine yielded the best performance. The lack of sufficient training data for the weather alerts and the less repetitive nature of the material may account for these results.

Many SMT systems can be framed in terms of decision theory (Ferrer et al. 08) in which, a translator $c$ is seen as a classifier mapping the set of source sentences $X$ into the set of target ones $Y$. The goal is to find the classifier which minimizes the conditional risk for each observation $x$:

$$c(x) = \arg\min_{y \in Y} \sum_{y' \in Y} l(y \mid x, y') p(y' \mid x)$$

where $l(y|x,y')$ is the cost incurred by classifying $x$ as $y$, while the true class is $y'$. Nowadays, the most popular decision rule corresponds to a family of cost functions:

$$\varepsilon(x,y) = p(y \mid x)^{-1} \times \prod_{i=1}^{M} f_m(x,y)^{\lambda_i}$$

which relies on *M* feature functions (already trained statistical models) of the form:

$$f_i(x,y) = \exp\left[h_i(x,y)\right]$$

each one, being weighted by a coefficient $\lambda_i$. The values of the coefficients are typically adjusted on a development bitext by applying an iterative procedure, which optimizes an automatic measure of translation quality. The decision rule that follows is:

$$c(x) = \arg\min_{y \in Y} \prod_{i=1}^{M} f_i(x,y)^{\lambda_i} = \arg\min_{y \in Y} \sum_{i=1}^{M} \lambda_i h_i(x,y)$$

The language model is an essential feature function of a translation system since it controls the fluency of the translations produced. Typically, we use a so-called *n*-gram language model, that is, an *n-1* order markovian approximation of the likelihood of a sequence of *m* tokens $y_1^m = y_1,...,y_m$.

$$p(y_1^m) = \prod_{i=1}^{m} p(y_i \mid y_{i-n+1}^{i-1})$$

Naturally, the most important feature function is what we call a transfer model or so-called phrase table, that is, a huge table of pairs of source and target phrases that are translation of each other, each one associated to a score. Such a model is typically acquired by a heuristic procedure, which capitalizes on a word-alignment procedure (Koehn et al., 2001). Brown et al. (1993) presented a series of models for which the alignment can be derived entirely automatically from a bitext. In a nutshell, they introduced the notion of alignment between two sentences $x_1^n$ and $y_1^n$, which is an application from [1,*m*] into [0,*n*] (the null position is given to each target word which do not have a direct association to a token in *x*): $a_{i \in [0,n]} \forall i \in [1,m]$. This allows the decomposition of the joint distribution into so-called alignment distributions $p_a$ (very often a 1-order markovian process as shown), and lexical distributions $p_l$:

$$p(y_1^m, a_1^m \mid x_1^n) = p(m \mid n) \prod_{i=1}^{n} p_a(a_i \mid a_{i-1}) p_l(y_i \mid x_{a_i})$$

The alignment $a_1^m$ being unknown, an iterative maximum-likelihood procedure (Expectation-Maximization) is applied for adjusting the parameters of the joint distribution over a training bitext. In such a setting, computing the most likely alignment $\arg\max_a p(a_1^m \mid y_1^m, x_1^n)$ is equivalent to run the Viterbi algorithm.

In this project, we will develop a new translation engine by first reconstructing a phrase-based engine from scratch, using specific knowledge from EC in order to adequately select the training material. Since MT is a fast evolving scientific field, the latest technology will be integrated and improvements upon (Langlais et al, 2005) are to be expected. In particular, specific models will be trained for translating specific types of weather warnings. Second, we will design a warning specific translation memory as a front end to the translation device. Finally, we will apply an approach for improving the overall system: error driven learning (Brill, 1995), which has been shown to be effective in NLP tasks such as tagging. This technique, which can be seen as a post-processing strategy, is conducting a greedy-

search over a huge space of contextual transformation rules, which allow the modification of parts of a translation. The search is guided toward reducing recurrent errors over a development set.

Although French and English are the two official languages of Canada, there is now a strong incentive to publish environmental information in Inuktitut, the language of Nunavut, particularly given Canada's determination to affirm its sovereignty over the Artic. Our team already has some experience in the processing of Inuktitut (Langlais, Gotti and Cao, 2005), and we know this presents a big challenge, especially because there is not much data in machine-readable format.

EC has already developed a prototype system, which provides automated voice broadcasting of some weather forecasts in Inuktitut. In a nutshell, the English words of the meteorological domain have been translated manually into Inuktitut; a pronunciation has been attached to each Inuktitut word, and a speech synthesizer concatenates on the fly the message that is delivered to the user. In our project, we will analyze ways of improving this process. For reasons of simplicity, we will concentrate on text only. We will make use of external resources such as the Canadian Inutktitut/English Hansards[1] (Martin et al. 2003) and morphological analyzers (Johnson et Martin 03) in order to help acquire language models that can be used to smooth the sequence of words currently synthesized by the system.

Integrating morphology into a translation engine is still a key-challenge of SMT (Toutanova et al., 2008). At RALI, we acquired a solid expertise in analogical learning (AL), a lazy learning approach which can be seen as an unsupervised way of mapping a sequence of source symbols (characters our words in our case) into a sequence of target ones. We studied several applications of AL to SMT (Langlais, Patry, 2007; Langlais et al.; 2009). In particular, we showed how a combination of both approaches could improve the quality of an engine translating into a morphologically rich language such as Finish, which shares with Inuktitut a strong compositional property.

## Relevance to the non-academic world

Timely weather and environmental information is important for decision-making related to the health and safety of citizens and protection of property. This kind of critical, real-time information is used by Emergency Measure Organizations (EMOs), as well as the transportation, agriculture, forestry, fishery, recreation and tourism industries, the military and the public. Each application has a slightly different focus, meaning that different users of weather reports are interested in a different set of details about a forecast. Tailoring individual reports manually can be very time-consuming and frustrating considering the short *shelf life* of a weather report.

Severe weather situations evolve extremely rapidly and although recent numerical models can produce very short-range forecasts, called *now-casts*, these require a considerable amount of manual intervention. Any automated (or even semi-automated) technique for generating them would be welcome. This project will develop specific data-mining techniques for dealing with the various types of weather information that are routinely produced in vast quantities by Environment Canada.

Environment Canada is also required to deliver forecast products in at least the two official languages. This project will adopt a *multilingual* perspective in the context of new information dissemination techniques in order to better deliver precise and up to date environmental information to all Canadians.

---

[1] Available at http://www.inuktitutcomputing.ca

## References

J. Andrés-Ferrer, D. Ortiz-Martínez, I. García-Varea and F. Casacuberta (2008), On the use of different loss functions in statistical pattern recognition applied to machine translation, Pattern Recognition Letters, 29-8, 1072--1081.

E. Brill, Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics* 21, 4, 543-565, Dec 2005.

N. Cressie, *Statistics for Spatial Data*. Wiley, 1993.

M. Fasciano and G. Lapalme. Intentions in the coordinated generation of graphics and text from tabular data. *Knowledge and Information Systems*, vol. 2, number. 3, p. 310-339, Aug 2000.

E. Goldberg, N. Driedger and R. Kittredge, Using Natural-Language Processing to Produce Weather Forecasts. *IEEE Expert: Intelligent Systems and Their Applications* 9, 2, 45-53, Apr 1994.

H. Johnson and J. Martin. Unsupervised learning of morphology for English and Inuktitut. In *Proceedings of the 2003 Conference of the North American Chapter of the Association For Computational Linguistics on Human Language Technology* - Vol 2, p. 43-45, 2003.

P. Koehn, F. J. Och and D. Marcu (2003), Statistical Phrase-Based Translation, HLT 2003, 127-133

P. Langlais, S. Gandrabur, T. Leplus and G. Lapalme. The Long-Term Forecast for Weather Bulletin Translation. *Machine Translation*, vol. 19, number. 1, p. 83-112, Mar 2005

P. Langlais, F. Gotti and G. Cao. NUKTI: English-Inuktitut Word-Alignment System Description. *2nd ACL workshop on Building and Using Parallel Texts: Data Driven and Beyond*, University of Michigan, Ann Arbor, p. 75--78, Jan 2005

P. Langlais and A. Patry, Translating Unknown Words by Analogical Learning, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p 877-886.

P. Langlais, F. Yvon and P. Zweigenbaum (2009), Improvements in Analogical Learning: Application to Translating multi-Terms of the Medical Domain, EACL 2009, 487-495.

J. Martin, H. Johnson, B. Farley and A. McLachlan. Aligning and Using an English-Inuktitut Parallel Corpus. *Proceedings of the HLT-NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 115–118, 2003.

A. Y. Ng, M. I. Jordan, and Y. Weiss, On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems (NIPS)* 14, 2002.

C.E. Rasmussen and C.K.I Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

E. Reiter and R. Dale (2000). *Building Natural-Language Generation Systems*. Cambridge University Press

E. Reiter, S. Sripada, J. Hunter, J. Yu, I. Davy, Choosing Words in Computer-Generated Weather Forecasts. *Artificial Intelligence,* 167:137-169, 2005.

M.L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

K. Toutanova, H. Suzuki and A. Ruopp (2008), Applying Morphology Generation Models to Machine Translation, ACL 2008, 514-522.

Pascal Vincent and Yoshua Bengio, Manifold Parzen Windows. *Advances in Neural Information Processing Systems (NIPS)* 15, 2003.

P. Vincent, H. Larochelle, Y. Bengio, P.A. Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. *Proceedings of the 25 International Conference on Machine Learning (ICML'2008)*, pages 1096-1103, 2008.

J. Yu, E. Reiter, J. Hunter, C. Mellish, Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13:25-49, 2007

# PROJECT LEADER AND PARTICIPANTS

## Project Leader

Guy Lapalme, Professor
RALI - Département d'informatique et de recherché opérationnelle
Université de Montréal
CP. 6128, Succ. Centre-Ville
Montréal, Québec
Canada, H3C 3J7

Guy Lapalme has been working in Natural Language Processing for more than 25 years focusing on the areas of spelling and grammar checking, text generation, summarization and machine-aided translation. Since 1997, he has led the RALI research laboratory (RALI=Recherche Appliquée en Linguistique Informatique). The RALI includes 3 faculty members, 2 research staff and 15 graduate students and post-doctoral fellows. It is the largest university based group devoted to Natural Language Processing in Canada. Guy Lapalme has a strong and continuous publication record and he is the recipient of an important annual Discovery Grant from NSERC (see NSERC Form 100). He also holds team grants from FQRNT and SHRC. He is currently leader of an NSERC CRD grant with an industrial partner. He has long experience in collaborating with industry in the implementation of practical Natural Language Processing systems.

## Project Participants

Philippe Langlais, Associate Professor
RALI - Département d'informatique et de recherché opérationnelle
Université de Montréal
CP. 6128, Succ. Centre-Ville
Montréal, Québec
Canada, H3C 3J7

Philippe has been working in the area of Natural Language Processing since 1995, when he started in the field of speech recognition. He joined the RALI in 1998, where his scientific interests focus on Machine translation and other multilingual applications. He too has had fruitful collaborations with industry, including a Precarn project on translation memories. He is a member of the board of two scientific journals, *TAL (Traitement Automatique des Langues)* and *Machine Translation,* and acts as a reviewer for many international conferences.

Pascal Vincent, Assistant Professor
LISA - Département d'informatique et de recherché opérationnelle
Université de Montréal
CP. 6128, Succ Centre-Ville
Montréal, Québec
Canada, H3C 3J7

Pascal Vincent's area of expertise is statistical machine learning algorithms and their application to data-mining. He has been working in this research field since 1996, both in academia (Ph.D. at Université de Montréal) and in the industry in well-known research laboratories of Lucent Technologies Bell Labs, AT&T Labs Research and Microsoft Research. He is co-founder of ApSTAT Technologies, a company

that, since 2002, delivers custom data-mining solutions for the insurance and finance industries. In 2006 he became professor at Université de Montréal, within the LISA laboratory (Laboratoire d'Informatique des Systèmes Adaptatifs), where he started a strong fundamental research program in his field, supported by an NSERC Discovery Grant, as well as a participation, with the 2 other professors of the LISA lab, in an NSERC strategic project grant. He is also part of the MITACS research project "Statistical learning of complex data with complex distributions" (lead by Yoshua Bengio). Pascal Vincent currently directs or co-directs 1 Ph.D. student and 3 Master's students.

# INDUSTRIAL COLLABORATION

Michel Jean, Director,
Meteorological Service of Canada - Quebec Operations
Environment Canada
E-mail: michel.jean@ec.gc.ca
Telephone: (514) 283-1600
Fax: (514) 283-1604

Environment Canada has committed 25K$/year, see confirmation letter of Michel Jean, for three years to this project starting from Sept 2008. The first 25K$, to be received as a Grant and Contribution, will be used before the MITACS Seed Project to do a literature survey of the area, to better define the problem, to prepare the source data and corpora that will be used in the project described in this proposal. The two next instalments of 25K$ will be available as of April 1st 2009 and 2010 and will coincide with the two years of the MITACS Seed Project.

Environment Canada will be involved in all stages of the project: from the definition and selection of the specific problems to the application of the research. Regular communication between contacts at EC will take place and some students are expected to work on the EC premises to make sure that the projects will be applicable in the organization. Moreover we plan to organize semi-annual meetings between all members of EC and the university researchers involved in the project.

A research agreement will be signed between Environment Canada and Université de Montréal that will allow the joint publication of the research results developed in this project.

# DEVELOPMENT OF HIGHLY QUALIFIED PERSONNEL

All phases of this project will involve graduate students who will learn innovative methods in natural language processing, data mining and visualisation and how to integrate them in an applied context. As shown in the budget, most of the money will be used for funding three graduate students (2 M.Sc and 1 Ph.D) to work on the research aspects of the proposal.

Each student will be supervised by a professor, along with the assistance of a research professional who will ensure that the data and the programs are available and will take care that the programs developed by the students can be smoothly integrated in the computing systems of Environment Canada.

# NETWORKING

For more than 10 years, RALI has organized a weekly seminar series[2], which allows students to become aware of recent developments in natural language processing. Students are also regularly invited to present their own work. These seminars are organized jointly with the Observatoire de la linguistique Sens-Texte (OLST) in the Linguistics and Translation department of the Université de Montréal. Seminars alternate weekly between the linguistics and computer science departments.

LISA also organizes a series of weekly seminars in the areas of data-mining and machine-learning, called *Séminaires UdeM-McGill-MITACS d'Apprentissage Automatique*[3].

These seminar series are advertised to both groups and researchers and students from one group regularly attend seminars of the other. This project would provide additional impetus for linking the two groups.

The Canadian Meteorological Centre at Environment Canada also organizes its own series of internal seminars, and researchers and students in this project will be invited to present the state of advancement of their projects there.

[2] The list of current and past seminars is available at http://rali.iro.umontreal.ca/Seminaires/info.html
[3] The list of current and past seminars is available at http://www.iro.umontreal.ca/article.php3?id_article=107

# BUDGET AND MANAGEMENT

The following table describes the revenues and expenses in the project.

| From | Sept-2008 | Apr-2009 | Apr-2010 | |
|---|---|---|---|---|
| To | Mar-2009 | Mar-2010 | Mar-2011 | |
| **Revenues** | | | | |
| Environment Canada | | | | |
| **Cash Contribution** | **25 000$** | **25 000$** | **25 000$** | |
| Inkind | 10 000$ | 10 000$ | 10 000$ | |
| **MITACS** | | **50 000$** | **50 000$** | |
| | | | | |
| **Expenses** | | | | |
| Students | 12 000$ | 51 000$ | 51 000$ | |
| PhD 1 | 12 000$ | 18 000$ | 18 000$ | 1: Information selection |
| MSc 1 | | 16 500$ | 16 500$ | 1: Information display |
| MSc 2 | | 16 500$ | 16 500$ | 2: Inuktitut SMT |
| | | | | |
| Professional | 13 000$ | 13 000$ | 13 000$ | Warning SMT + Coordination |
| | | | | |
| Equipment | | 3 000$ | | |
| | | | | |
| Travel | | 8 000$ | 11 000$ | |
| | | | | |
| **Total** | **25 000$** | **75 000$** | **75 000$** | |

The Sept 2008-Mar 2009 period will be financed by a contribution from Environment Canada that will be spent before the start of the current Seed Project. It will partially fund a PhD student (because the funds cover only 2/3 of the 2008-2009 academic year) who will conduct a literature search, gather initial corpora and data, and develop appropriate data-mining techniques before the arrival of the Master students in April 2009.

The description of the work of the students during the Seed Project was given in the first section above.

Pascal Vincent and Guy Lapalme will co-supervise the MSc 1 and the PhD projects as they involve important aspects of data mining combined with natural language processing. Philippe Langlais will supervise the MSc2 project dealing with the Statistical Machine Translation part.

The Professional (1/4 time in 2009 and 2010) is an experienced RALI research staff (Fabrizio Gotti) who will help set-up the project in the first months (1/3 time in 2008) and begin the development of the statistical machine translation system for weather warnings, He will also ensure the coordination of the work of the students and maintain regular contact with the partners at Environment Canada.

The equipment budget is for buying two high-end Linux workstations for two graduate students. The other graduate student and the professional will share the current facilities at the RALI and LISA.

Travel funds will be used for the cost for students and researchers (between $2 000 and $3 000 for each trip) to attend international conferences (AMTA, ACL, EMNLP) to present the work that they will have completed in the course of the year.

# RELATIONSHIP TO OTHER RESEARCH SUPPORT

Guy Lapalme and Philippe Langlais each hold a Discovery Grant and are co-investigators in a CRD grant from NSERC in collaboration with Terminotix, a firm that sells translation tools and services. Although these projects also involve natural language processing, they have no direct overlap with the work described in this proposal, which is targeted to the specific needs of Environment Canada.

Pascal Vincent holds a NSERC Discovery Grant and is part of an NSERC-STPSC grant, and a MITACS project grant. The research program and projects that these grants support focus on investigating fundamental issues in machine learning. They are not related to the present project proposal, which is targeted at a specific application of data-mining and natural language processing technologies to the data and problems of Environment Canada.

# ENVIRONMENTAL IMPACT CERTIFICATION

The work described in this proposal is limited to the information processing aspect of weather information dissemination. As such, it has is no direct environmental impact.

# CONFLICT OF INTEREST DECLARATIONS

The project leader and investigators have no involvement in the organization of Environment Canada or in any other organization doing business with Environment Canada.

# REFEREES

- Ehud Reiter [expertise: Natural language generation]
  Department of Computing Science
  University of Aberdeen
  King's College
  Aberdeen AB24 3UE
  Britain
  e.reiter@abdn.ac.uk

- Fred Popowich [expertise: Machine translation in an industrial context]
  Simon-Fraser University
  8888 University Drive
  Burnaby, B.C., V5A 1S6
  popowich@sfu.ca

- José Coch, Lingway [expertise: Generation of multilingual weather bulletins]
  Immeuble PARITALIE
  18, rue Pasteur
  94278 Le Kremlin Bicêtre cedex
  jose.coch@lingway.com

- Graeme Hirst [expertise: Natural language processing]
  Department of Computer Science
  University of Toronto
  Toronto, Ontario
  M5S 3G4
  gh@cs.toronto.edu

- Andy Way [expertise: Statistical machine translation]
  School of Computing
  Dublin City University
  Glasnevin, Dublin 9
  away@computing.dcu.ie

- Dr. Pierre Isabelle [expertise: machine translation and Inuktitut language processing]
  Technologies langagières interactives
  Institut de technologie de l'information du CNRC
  283, boulevard Alexandre-Taché
  Édifice CRTL, pièce F2-007
  Gatineau, QC J8X 3X7
  Pierre.Isabelle@cnrc-nrc.gc.ca