Introducing the Natural Language Generation of Text Weather Forecasts in the GFE

Article · July 2009

CITATIONS
3

READS
333

1 author:

Tennessee Leeuwenburg
Bureau of Meteorology
13 PUBLICATIONS **160** CITATIONS

SEE PROFILE

Introducing the Natural Language Generation of Text Weather Forecasts in the GFE

Tennessee Leeuwenburg

Centre for Australian Weather and Climate Research

Bureau of Meteorology

t.leeuwenburg@bom.gov.au

Introduction

The text formatters in the Graphical Forecast Editor (GFE) bring together the two disciplines of computational linguistics and weather science. The formatters are software modules, capable of pursuing goal-directed behaviour to generate English-language weather forecasts in a similar style to that used by weather forecasters in live operations.

Whereas much scientific research concentrates on the supply of good numerical forecast guidance, the work involved in producing text formatters concerns itself with how to represent the most important aspects of that guidance as text to end-users. Conveying the correct meaning in a piece of text can be critical to life (as in the case of warning forecasts). The text formatters must guarantee the full reporting of that information, in addition to being assessed according to the clarity and elegance of the forecast.

Computer users will be aware of how rigid any automated system can be. The text formatters have a complex engine driving them, which allows for high-level, language-directed goals to be captured. There are two vectors of work involved in producing excellent weather forecast text. The first relates to work on the fundamental capabilities of the text formatters. The second relates to configuring and specialising that fundamental capacity so that it delivers reports which are in line with the requirements of various stakeholder groups.

The formatters provide the means by which scientific rigor in weather description can be maintained. For the first time it is possible to be clear about exactly what conditions will trigger a warning condition, in a way which relates directly to the underlying guidance data. Removing the human from first-guess forecast generation means that precise meaning can be attached to the words used, rather than relying on the analogy or mental construct of the weather situation to arrive at the resulting description.

A Short Introduction to the GFE

Development of the GFE began in 1992 by the USA National Weather Service, first producing a full suite of digital products in 2003 (LeFebvre, Mathewson and Hansen 2003). It is operationally deployed across the USA, producing both graphical and text-based forecasts.

The GFE is made up of a graphical client and a data server, with the bulk of the functionality being initiated by a human forecaster. The client allows forecasters to not only view, but directly manipulate, NWP forecast guidance, other forecast guidance systems and observational and analysis data.

One major shift with the introduction of the GFE is that text and graphical forecast production systems are almost fully automated. The main work of the forecaster is producing a modified set of forecast data grids. If this were done in a fully ad-hoc way, the potential would exist for these grids to become quite implausible in physical terms. Unlike NWP modeling, conservation of energy and parameter consistency is not maintained, nor is it feasible to do so.

However, in order that grid manipulations (a) are reasonably easy to perform, (b) operate in a scientifically defensible manner to the extent possible and (c) support the range of operations required to create a set of forecast grids,

automated tools known as Smart Tools have been developed. The first Australian set of these, representing years of effort from CAWCR and NOAA, are a launching pad for further research and development. They open up avenues for research into algorithmic post-NWP mechanisms, such as capturing topographic effects and deriving weather parameters which are not explicitly modeled.

Once a set of forecast grids is thus produced, the forecast product generation is highly automated. (While graphical forecast products may be generated, the current research focus is on text forecasts.)

The text formatters, which do this work, operate with four components of evaluation:

1. Statistical sampling and data representation
2. Information processing
3. Language representation
4. Post-processing

These stages may be considered sequentially, but there is always feedback between the components when actually producing forecasts.

Work on the text formatters, then, falls into a variety of areas:

1. Algorithms and techniques in language processing
2. Studies into the effective use of language generally (such as appropriate grammar, choosing a good lexicon of terms, understanding what makes a sentence clear and comprehensible)
3. An examination of how forecasters choose to express concepts regarding future conditions, both to mimic this and also to identify any surmountable communication difficulties.
4. The gathering of requirements from a services perspective relating to what must be guaranteed when a forecast is produced (for example, the guarantee that forecast text will always mention peak wind speeds during warning conditions, or that District forecasts will always include thunderstorms if they are forecast for more than 5% of the area).
5. Ensuring that terminology is used in a consistent fashion (quantification of forecast terms used).
6. Expansion of the expressions used to incorporate probabilistic guidance and new forms of forecast.

This work involves a variety of tasks, varying from fundamental research, to engineering and information management. The algorithms and theories that describe how to go about information processing and language representation are the focus of much on-going academic activity. Unlike physical science, this is often less mathematical and more concept- or design-based. It is a less mature area of research and can involve much searching in the dark.

The examination of forecast terminology and consistency between forecasts has not previously had a large amount of resource directed towards it, partially due to the human factors involved. While there is a call for greater consistency of terminology, it has become clear during the GFE implementation that not all concepts of ‘consistent’ are equally applicable. The statistics which make a forecast term appear in one product but not another (e.g. a Town versus a District) can result in the appearance of inconsistency between written forecasts, even when the underlying rules are being applied in a consistent fashion. Negotiating the relationship between the rules which define the system and the requirements which are placed on the output of the system is too much a matter of experience and intuition.

The tacit knowledge held by a human forecaster is something which needs to be drawn out and considered, rather than a source of consistent advice regarding the appropriate behaviours to attempt to build in to the automated system.

While the drivers for GFE development are currently coming strongly from the IT discipline (that is, being driven by bug reports and feature requests) rather than science (verification results, experimentation and evaluation), it is expected that with the national rollout, and with increased certainty in the future of the GFE, that progress in the fundamental capabilities of the text generation system will come from applying research techniques as much by ad-hoc problem-solving. Getting the most from the partnership with NOAA will also be important.

Text Generation and Computational Linguistics

There are a number of key papers within the domain of Computation Linguistics which are useful for placing the GFE text formatters into academic context. One such paper is presented by De Smetd et al., (1996). This impeccably-researched paper critically analyses a number of existing Natural-Language Generation (NLG) systems and sketches some new directions for further research. In their language, the GFE formatters exhibit features of both interactive (feedback) systems, blackboard systems and revision-based systems. These architectural types refer to the way in which the relevant systems proceed through two well-recognised stages: text or document planning and surface realization or lexicalization – or in layman's terms, what to say and how to say it (p3).

As acknowledged, these stages are not entirely independent. What you want to communicate is influenced by what you are able to say. Some concepts are remarkably hard to describe, while others can be expressed in just a few words. De Smedt, Horacek and Zock also state that few existing systems are really very capable:

"*In the area of text planning, researchers have identified more problems and limitations of existing theories than they have provided new solutions. With regard to lexicalisation, only a few researchers have made proposals that go beyond one-to-one mappings between concepts and words…*"

The GFE formatters have a number of features which place them at the more complex, more sophisticated end of the spectrum. In other respects, they are probably a little behind the game. Where they excel is in document planning and the way in which document planning and lexicalization dynamically interact with one another. In terms of shortcomings, more work is necessary to extend the number of grammatical constructs they can support. It needs to be easier to quickly implement new phrases.

A fertile area for new research is partial specification. Frequently, systems are put together without a distinction between must-goals and should-goals. An example of a must-goal would be to always mention the timing of when wind speeds increase past a warning threshold. An example of a should-goal is to describe wind changes in 5-knot increments, but this may come second to other should-goals relating to ideal sentence length.

The GFE has no inherent concept of a must-goal in terms of explicitly listing and separating the two goal types. Indeed, there is no explicit separation of goal-directed and rule-based processing components. However, both goal-directed behaviour and dynamic goal conditions have been implemented into the basic structure of the formatters.

It is also enlightening to compare the GFE text formatters to its close cousins. SumTime-Mousam (STM) is an NLG system for wind forecasts (Reiter et al., 2005). Further, Reiter is one of only a few authors to include a proper evaluation of the relative performance of the automatically-produced text as compared with human-authored text. Information on the entire scope and capability of the system is limited, with the research reports and available web information concentrating only on the wind-phrase generation logic. FOG is another NLG system for Atlantic Marine forecasts by the Atmospheric Environment Service of Environment Canada (Goldberg, et al., 1994).

The research findings are very much relevant to the development of the GFE formatters, both in terms of the computer science aspects and in terms of the lexicon-design and requirements-negotiation processes. We will first look at the implications for system design, before addressing issues of word choice and requirements later in this article.

STM is less complex in many respects than the GFE formatters, but with some particular advantages over the GFE formatters in terms of an expanded lexicon of verb terms (which it uses particularly effectively). They key difference between the two systems is the capacity of the GFE to react to the complexity of the meteorological situation and adjust its should-goals appropriately.

STM will seek to represent every detail of the situation entered as a data input, potentially resulting in an overly-lengthy phrase. To combat this, STM also employs the use of acronyms to describe compass-point directions. By contrast, the GFE formatters have to include a great deal of situation-recognition logic in order to distil from

raw statistics what is significant about a weather situation and how to summarise that effectively while still communicating the salient points. STM appears to have a two-level detail response capacity, whereas the GFE has up to eight levels of detail with independent settings for strong or light prevailing conditions.

Architecturally, STM appears to explicitly represent various NLG stages: document planning, microplanning (lexicalization) and aggregation. Aggregation is used in the STM research to mean the distillation of information among sentences.

Compared with STM, the GFE formatters have greater responsibility for distilling information from raw statistics and more capacity to capture high-level goals, but have a slightly less capable lexicalization system as currently deployed. (That is to say, it is not so much the system which is limited, as the number of words currently used to describe changes in conditions.)

Text Generation and Meteorology

The main links between text generation and meteorology are data processing and information representation. To date, the key performance indicator of the GFE has been the verification results of forecast data grids produced by the system. Examples of such work include Stern (2007). In terms of general, written-forecast metrics, available research concentrates on the statistical accuracy of the forecasts rather than on the semantic clarity of the forecasts (Toth et al., 2006).

It is believed that this is in large part due to the lack of a clear quantitative method for assessing the semantic clarity of a piece of written text. As such, assessment has taken the form of qualitative analysis, survey results and intuitive response.

All of this makes verification statistics of text forecasts extremely hard and expensive to produce. Evaluation is a painstaking manual process. It is possible to glimpse how more objective metrics of forecast text might be derived, such as performing an analysis of grammatical complexity, but such things are not achievable now.

Further, the quantitative analysis of operational forecasts has tended to be used for post-development, forecaster evaluation. That is to say, the verification of guidance systems has been "over here", while the analysis of forecaster performance has been "over there".

The challenge, then, is to discover the links between meteorology and text generation, and what each discipline can learn from the other.

The text formatters, as a means of mechanically producing text forecasts (potentially from raw model guidance), significantly narrow the distance from here to there. In so doing, they provide meteorological researchers with the means to much more quickly influence the worded text forecasts.

One example of this in the GFE is the use of probabilistic information. While the GFE currently only makes use of such information as input into derived weather grids (which store non-probablistic coverages such as "scattered"), this is likely to change. As more probabilistic guidance becomes available through ensemble modeling, the capacity exists to very quickly reflect that additional information in the forecast words. Previously, such a change would have required significant forecaster training.

This potentially brings in a much closer connection between forecast expression and meteorological knowledge. Scientists who are working on novel systems have the potential to have a role in determining how their information could be integrated into the final forecast products. Information also flows the other way – gaps in what can be described in the text forecasts could potentially highlight gaps in the guidance systems.

Word Choice and Expression

It is the issue of word choice and grammar which have dominated discussions when building the formatters in the GFE. I used to describe these issues as being moving the tip of an iceberg. Changing the words also involves change to a large, hidden body of Python code which does not move as fast as new expressions can be imagined. Another, equally valid metaphor is that of genotype and phenotype. The requirements are all on the phenotype: what the forecasts should look like, while the work is all on the genotype: how the formatters are coded and configured. There is an enormous amount of semantic awareness held

by both researchers and forecasters alike, much of which is consciously accessible, and much of which is only tacit.

This introduces several complicating factors, not the least of which is the time-consuming process of talking to people about their tacit knowledge, in order to try to systematize and capture that knowledge. The major issues are essentially due to (a) the unforeseeability of exactly what new language will emerge from implementing a new rule, and (b) differences between important stakeholder groups regarding what is the appropriate way to describe a given weather situation.

Before the advent of the GFE, individual differences were accommodated by simply allowing small differences to be reflected in the official forecasts – indeed those differences would often pass entirely unnoticed. It was not just possible, but actually the case, that one forecaster might use a term such as "chance" to describe a situation, where the next person would use "isolated". The problems were more serious when individual instincts regarding whether conditions were serious enough to warrant issuing a warning came into play.

Nobody questions whether forecasters act with skill or to the best of their ability, but individual differences were generally not forced into alignment. Further, forecasters were given little opportunity to learn whether their forecasts were resonating with the public or not. Such feedback as was available came as the result of intermittent, ad-hoc and unfocussed feedback surveys rather than being an ongoing part of their jobs.

These differences of opinion are also reflected in some ways within the individuals involved in setting direction and policy. Every person involved in the GFE text formatter process, including the developers, managers, reviewers and forecasters has taken part in a discussion about what is the best way to talk to the public about the weather. The gulf between a naïve understanding of the weather and a meteorologist's understanding of the weather is sometimes very large.

For better or for worse, the GFE formatters support a deterministic text generation approach and individual preferences are not supported. Since the formatters produce a consistent style, each individual forecaster now has to sit down in front of hundreds of pre-generated forecasters which may or may not sit comfortably with their natural language style. While it has been made clear that the forecasters are not responsible for making stylistic changes to the GFE forecasts, nonetheless the forecasters are accustomed to taking full responsibility for the forecasts issued on their watch and as such often have strong opinions regarding the clarity and style of the language used (numerical forecast accuracy aside).

To make this really clear, it appears as though most meteorologically aware individuals have similar, but non-identical, mental analogies by which they think (and naturally, talk) about the weather. The public does not always share those analogies, which can result in confusion and disagreement. A similar issue might arise between a sound engineer and concert pianist – while both are concerned with music, they will often use entirely different terms and concepts to think and talk about it.

It is worth going through some examples in order to really see exactly what kind of problems are presented, and how significant they can be. We will first look at an example of this problem from the perspective of the forecasters being confronted with some actual examples of formatter output. Consider the following description of the wind:

"Winds northeasterly 15 knots tending northwest to southwesterly around midday"

This example complies fully with the requirements that resulted from a long period of discussion and evaluation of formatter performance prior to the GFE go-live date, yet it has emerged as a significant issue since that time. To naïve eyes, there may be little to distinguish this sentence from the preferred alternative:

"*Winds northeast to northwesterly 15 knots tending southwesterly during the afternoon"*

The difference lies in what the forecasters feel is not being communicated in the first example, namely that there is a frontal passage moving through which results in a generally southwesterly wind flow. Neither sentence is inconsistent with the data but one assembly of the sentence fits with the forecasters' mental models of the evolving weather situation.

There are many more such examples. In some cases, it is clear that the forecasters' instincts are correct. In a few cases, the forecasters have become accustomed to using expressions which are outright unclear. It is in the middle that most of the debate occurs. There is also a tension between using simple, easy-to-understand forecasts and elegant expression. This is also where a lack of capability in the generation system limits what can be produced, and why fundamental linguistics research is still necessary. For example, the use of overly simple language can result in highly repetitive forecasts which are actually more difficult to read and comprehend.

The work by Reiter et al (2005) goes to some lengths to determine how effective automatically generated text is, compared against human-authored text and against human post-edited, automatically pre-generated text. Their findings are fairly clear – that human post-edited forecast text is preferred by readers. Their analysis of the reasons for this is in line with the opinions of the GFE formatter development team.

The major advantage of automatically generated text is a consistency which is not possible given the individual preferences of human forecasters. However, the human capacity for expression is significantly greater than that of automated systems. Presenting a first-draft forecast which uses consistent semantics generally means that a human forecaster will adopt the same semantics for the final forecast. They are able to improve upon the language of the automated system without compromising semantic consistency.

Reiter et al (2005) also delve into which particular aspects of forecast language are handled better by the automated system as opposed to human authoring. They find that in human-authored forecasts, some kind of term (for example verb choice), are based almost entirely on the data at hand, while others (such as times of day) are described poorly and inconsistently.

Conclusions

The text formatters provide a step forward in the number of forecasts which can be produced. By producing forecast grids, the forecasters enable the automated system to produce numerically-consistent forecasts for a larger number of locations than current forecaster resources permit.

A greater degree of consistency of the terms used to describe the weather is also achieved, with all terms used now having a consistent and clear link to underlying atmospheric conditions.

The greatest challenges to the formatters lie in their ability to extract the salient information from the statistics, then render that information into clear and elegant language, taking into account the desired length and level of detail for each forecast.

References

De Smedt, K. Horacek, H. and Zock, M. 1996, *Architectures for Natural Language Generation: Problems and Perspectives*, in *Trends in Natural Language Generation: An Artificial Intelligence Perspective*. Berlin, Germany, Springer-Verlag pp. 17-46.

Goldberg E, Driedger N and Kittredge R I 1994, *Using Natural-Language Processing to Produce Weather Forecasts,* IEEE Intelligent Systems, Vol. 9, No. 2, pp. 45-53.

LeFebvre T, Mathewson M and Hansen T 2003, *The Rapid Prototype Project*, 19th International Conference on Interactive Information Processing Systems.

Reiter E, Sripada S, Hunter J, Yu J and Davy I 2005, *Choosing Words in Computer-Generated Weather Forecasts*, Artificial Intelligence, Vol. 167, Issues 1-2 pp. 137-169.

Stern H 2007, *Increasing Forecast Accuracy by Mechanically Combining Human and Automated Predictions using a Knowledge Based System*, 23rd Conference on Interactive Information and Processing Systems.

Toth Z, Talagrand O, Zhu Y 2006, The attributes of forecast systems: a general framework for the evaluation and calibration of weather forecasts, from Predictability of Weather and Climate by Tim Palmer and Renate Hagedorn, Cambridge University Press.