AEMIX: Semantic Verification of Weather Forecasts on the Web

Angel-Luis Garrido[1], María G. Buey[1], Gema Muñoz[1] and José-Luis Casado-Rubio[2]
[1]*IIS Department, University of Zaragoza, Zaragoza, Spain*
[2]*Spanish Meteorological Service (AEMET), Madrid, Spain*
{*garrido, mgbuey, gmunoz*}*@unizar.es, jcasador@aemet.es*

Keywords: Weather Forecast, Information Extraction, Ontologies, Automatic Verification

Abstract: The main objectives of a meteorological service are the development, implementation and delivery of weather forecasts. Weather predictions are broadcasted to society through different channels, i.e. newspaper, television, radio, etc. Today, the use of the Web through personal computers and mobile devices stands out. The forecasts, which can be presented in numerical format, in charts, or in written natural language, have a certain margin of error. Providing automatic tools able to assess the precision of predictions allows to improve these forecasts, quantify the degree of success depending on certain variables (geographic areas, weather conditions, time of year, etc.), and focus future work on areas for improvement that increase such accuracy. Despite technological advances, the task of verifying forecasts written in natural language is still performed manually by people in many cases, which is expensive, time-consuming, and subjected to human errors. On the other hand, weather forecasts usually follow several conventions in both structure and use of language, which, while not completely formal, can be exploited to increase the quality of the verification. In this paper, we describe a methodology to quantify the accuracy of weather forecasts posted on the Web and based on natural language. This work obtains relevant information from weather forecasts by using ontologies to capture and take advantage of the structure and language conventions. This approach is implemented in a framework that allows to address different types of predictions with minimal effort. Experimental results with real data are promising, and most importantly, they allow direct use in a real meteorological service.

1 INTRODUCTION

Humankind has lived for years watchful to weather and has tried to know how to predict atmospheric conditions for a given location. Today we continue giving to it a major importance, since many human activities have a great relationship with meteorology: agriculture, transports, or sport events. Any outdoor event like a concert or parade, or just planning a leisure weekend or holidays are also situations where a particular weather forecast can have a great influence at the time of making decisions.

Therefore, an important aspect for both companies and particulars is the broadcast of weather forecasts. The first daily forecasts were published in newspapers in the 19th century, and the first radio weather forecasts were made in the beginning of the 20th century. They were followed immediately by television broadcasts. Therefore, as expected, the popularization of the Internet in the 90s and the increase of web pages created a successful new form of dissemination of weather forecasts.

In each country, the meteorological services created their own websites where they publish predictions. Some examples can be appreciated in Figure 1. Soon, there were many commercial sites that also provided similar information. Today, with the growth of mobile devices, hundreds of applications for consulting weather forecasts have appeared for the different mobile operating systems. These mobile applications allow us to practically check in real time the evolution of the weather, and even tell us when an adverse weather event is going to happen.

Forecasters make their own interpretation of the mathematical models and create graphics, maps, and texts in natural language to explain the weather conditions of the atmosphere which may occur in the next few hours or days. This interpretation is what we see every day in our personal computers and our mobile devices through Internet technologies. An important task is to check the weather forecasts contrasting them to the data coming from actual observations. This is an interesting work, which can provide useful information to meteorological services, such as:

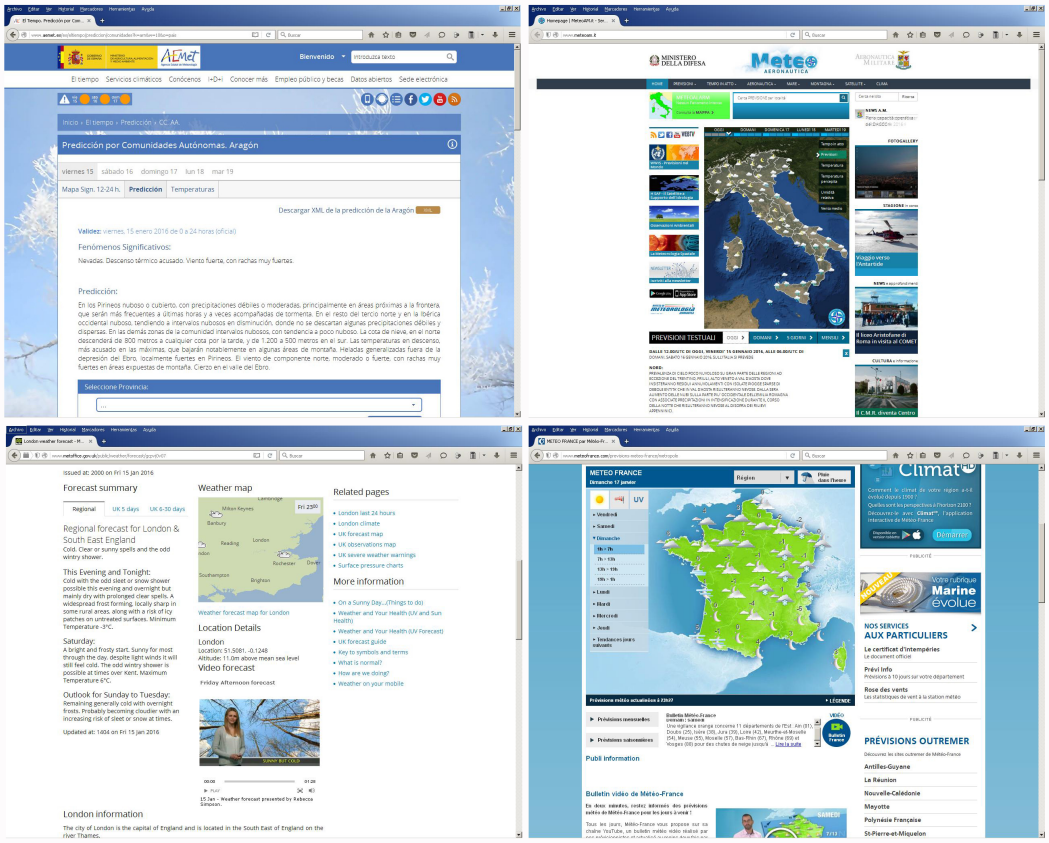

Figure 1: Meteorological Service's web pages offering wording weather forecasts.

- Empirical testing of the accuracy of forecasts.
- Verification of predictions made by automatic systems compared to those manually performed by humans.
- Detection of frequent and significant biases in predictions. Once they are identified, they can be analyzed in order to try possible solutions.

The problem emerges when a forecast has been published using natural language, because the verification is not trivial. It is necessary to transform the worded forecast into verifiable numerical data, and then compare these data to the actual observations from different meteorological stations situated in the forecast area. These texts usually predict different atmospheric variables: temperature, precipitation, cloudiness, among others. Accordingly, the type of data that is necessary to extract from the forecast varies. Furthermore, the content structure of predictions can be quite different, depending on the person (or the system software) who has written the forecast. Fortunately, it is common to have a writing style guide that helps to standardize weather forecast sentences. This problem has been already studied from a meteorological point of view during the last years (Mason, 1982; Murphy and Winkler, 1992; Jolliffe and Stephenson, 2012).

Hence, we propose a semantic approach for the development of tools to identify specific data (for example, atmospheric variables and their properties) from weather forecasts published on the Web in text format in order to verify them. The process is guided by the information stored into a special type of ontology that contains the knowledge about the structure and the language use of the weather forecast, as well as references to pertinent extracting mechanisms. Finally, we also present an implementation of our approach, the AEMIX Project[1], which have been developed in collaboration with the Spanish Meteorological Service (AEMET). Figure 2 shows some captures of the software.

This work is structured as follows. Section 2 analyzes the state of the art. Section 3 explains the different steps of the proposed system. Section 4 studies the preliminary results. Finally, we conclude the paper in Section 5.

[1] AEMIX stands for *AEMet Information eXtraction.*

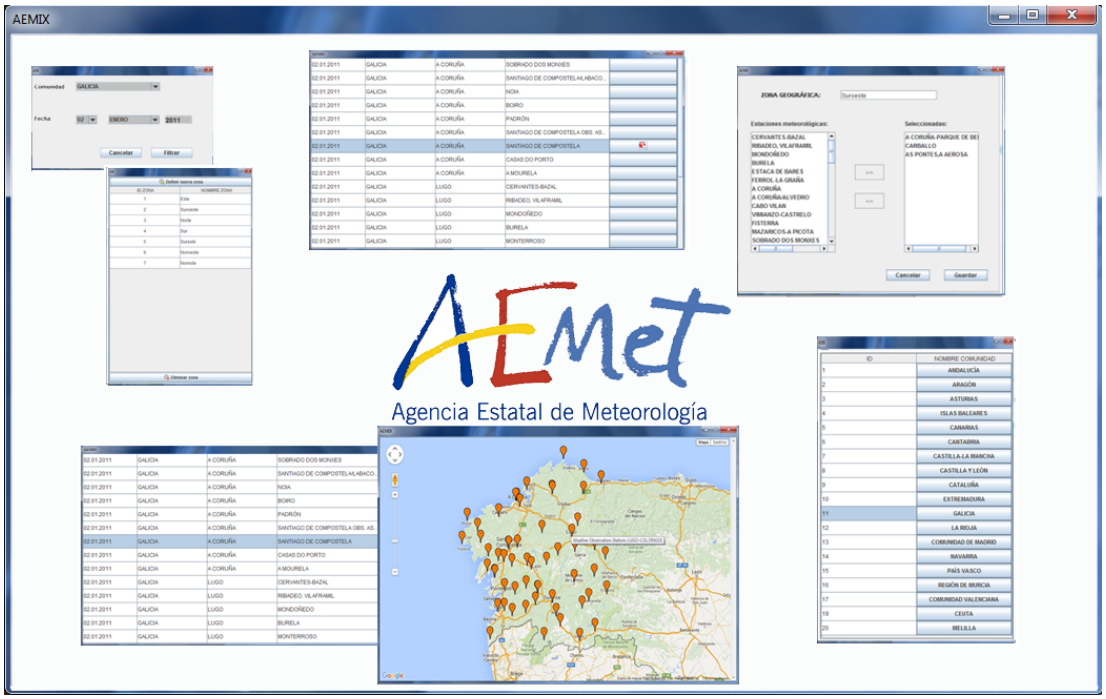

Figure 2: Some screenshots of the AEMIX software.

2 RELATED WORK

The Semantic Web (Horrocks, 2008) is an extension of the Web through standards by the World Wide Web Consortium (W3C), and it promotes common data formats and exchange protocols on the Web to provide a common framework that allows data to be shared and reused across application, enterprise, and community boundaries (Berners-Lee et al., 2001). The goal of the Semantic Web research is to allow the vast range of the web accessible information and services to be more effectively exploited by both humans and computer tools. In recent years, the influence of the Semantic Web in the design of computer programs, web services and applications in general is increasing, as can be found at (Aquin et al., 2008; Breslin et al., 2010; Garrido et al., 2011; Borobia et al., 2014; Buey et al., 2014; Bobed et al., 2016).

In order to determine and verify weather predictions, systems usually apply statistical calculations by joining the data which come from the predictions and observations. For instance, in (Murphy and Winkler, 1987), the authors proposed a general framework for forecast verification based on the joint distribution of forecasts and observations. Verifying the accuracy of weather predictions is an important task because the better the prediction system is, the easier it is to anticipate some critical situations (Mathiesen and Kleissl, 2011; He et al., 2009) .

Regarding weather issues, weather predictions are the result of numerical and statistical methods and techniques which try to anticipate the weather that is going to affect an area. They usually take into account the actual weather observations and their past trend. There exist some models such as the Numerical Weather Prediction (NWP), the Weather Research and Forecasting (WRF) or the COSMO model (Done et al., 2004; Baldauf et al., 2011) that allow to get weather forecasts, and also there are many approaches that focus on determining and verifying the actual effectiveness of them. This is an important task because their improvements anticipate some critical situations such as solar irradiance or flood alerts (Murphy and Winkler, 1987; Mathiesen and Kleissl, 2011; He et al., 2009).

The results of these predictions are given as a set of numerical data, so it might be difficult for the general public to understand them. In order to make these predictions understandable by everybody, it is necessary to transform them into a comprehensible format, e.g. to translate them into a natural language format. Many approaches have been developed to handle this task automatically (Goldberg et al., 1994; Reiter et al., 2005; Belz, 2008). Although there are many works dedicated to mathematical and statistical verification of weather forecasts, to the best of our knowledge, there are no works dedicated to the verification task of worded weather forecasts.

3 THE AEMIX SYSTEM

In this section, we describe our proposed system, called AEMIX, which identifies and extracts information contained in weather forecasts that are expressed in a natural language format and downloaded from the Web. This system aims at verifying that weather predictions match the real observation data in a specific date. Figure 3 depicts the complete process which has been divided into four stages.

The input of the system consists of a set of weather forecasts, and a spreadsheet with the equivalent real data from all the observation stations. After a preprocess, which stores all this information in a database, the system separates the plain text of the forecast into fragments according to the atmospheric variable described, and then each of these fragments is analyzed with the aid of an ontology.

AEMIX extracts relevant data from those fragments, and transforms them into equivalent numerical data. Once the system has converted the forecast into numerical data, it stores the information in a database, and, finally, verifies the results. This verification is the output of AEMIX.

We have used meteorological services style guides to identify in each weather forecast phrases referred to particular atmospheric variables. Each of these variables has an associated set of attributes and information about the data required to be found. For example, the data extracted from the precipitation might be its typology (drizzle, rain, snow, or hail), its adjectives and quantification (weak, moderate, heavy, very heavy, or torrential), and the temporal evolution of the atmospheric phenomenon (persistent, frequent, intermittent, etc.).

It is important to remark that there are certain attributes of atmospheric variables that can not be verified, because observations of them are sparse or nonexistent (for example, the presence of snow). The three most relevant variables to verify are the temperature, the precipitation and the wind.

As we mentioned before, the extraction process is guided by an ontology containing the knowledge about different meteorological variables, and how to identify and extract them in the text of a forecast. Following, we explain the different stages of the process.

As we can see in Figure 3, in the very first stage (”’Pre-processor”) we perform a number of different tasks which clean weather forecasts before the data extraction. These forecasts are downloaded from the web and they usually have a special format depending on the meteorological service. The main components in this stage are:

- *Geographical information:* Area where the weather forecast is valid. Namely, the name of a region or a country or even a continent.
- *Date and time:* Date and time of the writing of the forecast, either by human or by automatic systems.
- *Range or Type of forecast:* Period of the future which the prediction is valid for. Examples of these ranges can be one day, two days, a day and a half, etc.
- *Weather forecast text:* The plain text of the forecast. It usually needs to be cleaned, because it contains special characters used internally for communication. These characters are worthless data which hinder the work of extraction, so they should be filtered out.

This stage must be implemented with custom programming according to the data provided by the meteorological service web page. As soon as AEMIX has retrieved all the required information, the system stores it in a database in order to facilitate its access together with observation data from a standarized spreadsheet for the same date, also retrieved from the corporate website of the meteorological service.

In the second stage (”Text Analyzer”), AEMIX queries the ontology using a custom-made semantic interface to obtain the text structure that it expects according to the type of weather forecast (see Figure 3). With this information: 1) AEMIX analyzes the forecast text, 2) the system identifies the different possible meteorological variables, 3) it uses the most suitable method for cutting off the texts, also according to the information provided by the ontology, and 4), finally, for each variable, the system returns a set of tuples (`<Variable, Sentences>`) with the following information:

- *Variable*: Type of atmospheric variable: temperature, precipitation, storms, visibility, cloudiness or wind.
- *Sentences*: Group of sentences which that atmospheric variable refers to in a given forecast.

The third stage (”Data Extractor”) is in charge of extracting the data from a weather forecast and converting them to a numerical format (see Figure 3). The input is the set of tuples returned by the previous stage, including: 1) the name of the atmospheric variable, and 2) its related sentences in the text.

At this stage, AEMIX asks the ontology for the most suitable methods in order to extract the required data from the forecast. Therefore, the system owns

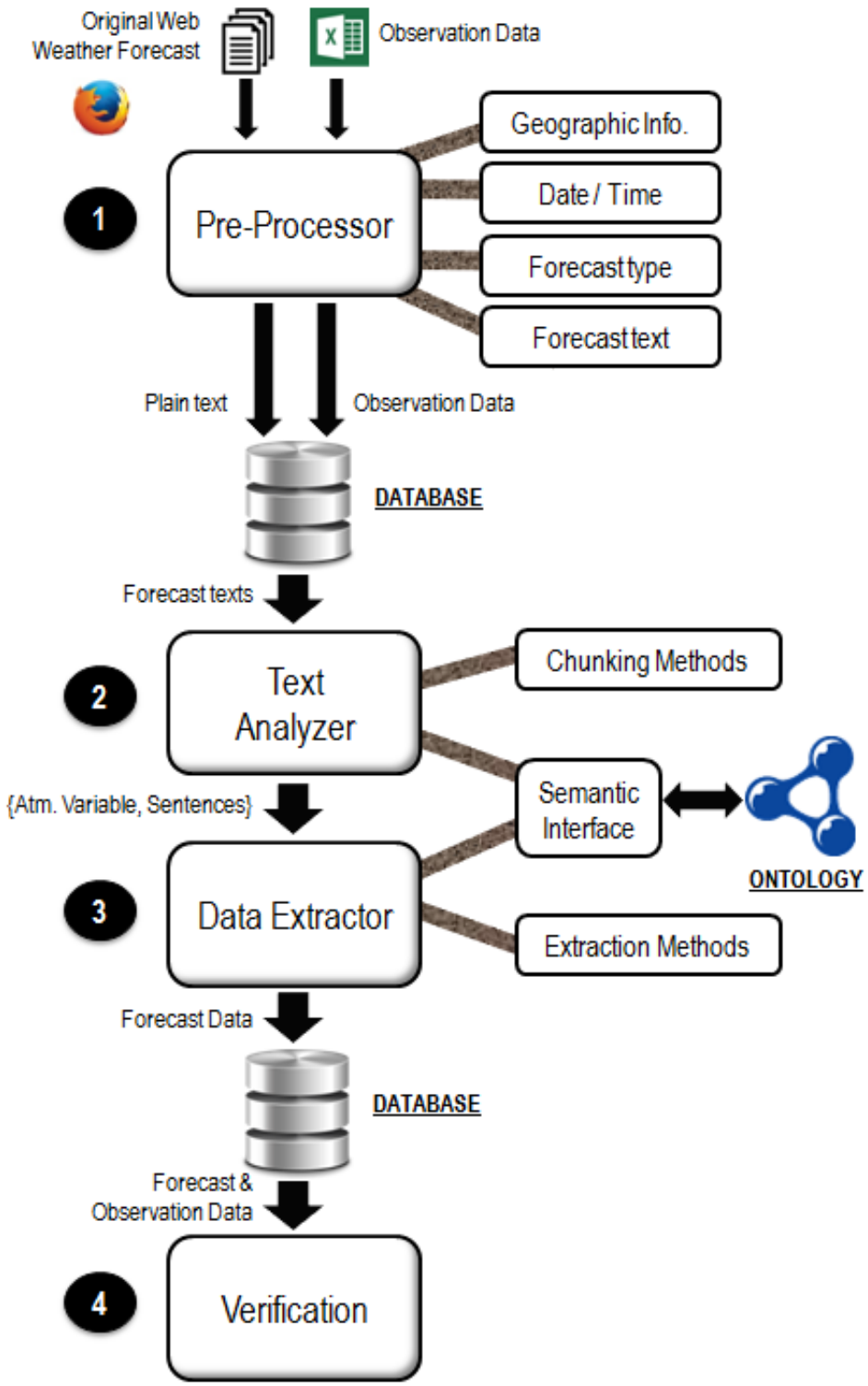

Figure 3: Stages of AEMIX system to extract numerical data from weather forecasts on the Web. on the Web. The system aims at verifying them, by using patterns embedded in an ontology that drives the extraction process.

several methods which can be applied on the input text. Using the knowledge of the atmospheric variable features depicted on the ontology, the system identifies the sentence format, and consequently, it uses regular expressions to recover the accurate data.

The output data of this stage are a set of tuples of elements (<Attribute, Value>) with the following information:

- *Attribute*: Particular characteristic of an atmospheric variable. As an example, if we are referring to temperature, we can mention "minimum", "maximum", or "frosts".
- *Value*: Possible attribute values. For example, in the case of wind, the possible values of the attribute *"direction"* are {"N"}, {"NE"}, {"SE"}, {"S"}, {"SW"}, {"W"} or {"NW"}.

As previously mentioned, there are certain attributes of an atmospheric variable that can not be verified. This is the case of frosts: there are not observations available referring to frosts. Therefore, on these cases, AEMIX returns a N/A (Not Applicable) value in order to know that there is no way to continue with the verification process.

When the system has recovered all the data, it transforms them into numerical information using a set of direct rules defined in the ontology. Whenever the attributes are geographical, the values are a set of geographical points defining the affected area, or a reserved word, like "Rest", which refers to the rest of the stations in the region included in the weather forecast, and not mentioned yet. When no location is indicated, it means that the forecast covers the whole region.

Therefore, the system consults the ontology through the semantic interface and obtains for each prediction a set of atmospheric variables identified in the forecast, and for each variable, AEMIX extracts a set of tuples (<Attribute, Value>). Finally, all these data are stored in the database.

In the fourth and last stage ("Verification"), we take for granted that we have stored in the database both observational data and forecast data (see Figure 3). These data are indexed to the valid date of the forecast and the observations, so the information for each atmospheric variable can be crossed and compared separately.

Regarding geographical data, if the prediction affects the entire region, the forecast is compared against observational data of all the meteorological stations. For example, when the forecast says that there will be weak showers, these will affect the entire region by default. In an attempt to simplify, we could say if 0 to 2 mm/h of rainfall are collected in most stations, the prediction will have been correct. But the verification process at meteorological level is much more complicated, and a detailed explanation is out of the scope of this work.

If the forecast affects only an area (defined by a longitude and latitude) using the attribute "location", only those observation stations situated within that area are taken into account to make the verification, and the predicted atmospheric variable is compared just against those observations. Therefore, the verification of the meteorological observations in an area will be done against those weather forecasts in the same area.

Hence, at the end of the process, the AEMIX system will house in its database all the detailed information about the two types of data (predictions and observations) related by date. Forecasters and meteorologists will be able to use this information to carry out data mining actions and perform any desired verifications.

4 PRELIMINARY RESULTS

To carry out the system testing, we used the information provided by the website of the aforementioned Spanish meteorological service (AEMET). We downloaded both the weather forecasts and observational data, and we have adapted both the ontology and extraction methods to AEMET's writing style guide. These extraction methods are based on symbolic pattern rules.

We used a sample of 2,828 weather forecasts corresponding to one year of predictions over Galicia region. The forecasts were classified into two types: FPSP75 and FPSP85, respectively corresponding to one-day or two-day forecasts. Corresponding observation data from 58 observation stations was also downloaded from the web, with the complete information about longitude and latitude. We obtained from the same source datasets of temperature and precipitation during the same year, for a total of 77,339 observation registers.

Regarding the geographical areas, we have identified the forecaster's linguistic uses to describe the main areas of the Galicia region, and then we introduced them in the AEMIX database, and next we related them with the corresponding geographical areas using an easy graphical interface.

Although the experiments are still ongoing, with our new enhancements, AEMIX achieved very good results (above 90% of correct results retrieved).

5 DISCUSSION

In this work, we have presented a new system, called AEMIX, able to extract information from worded weather forecasts downloaded from the Web. It aims at verifying their accuracy by comparing the forecasts against observation data from meteorological stations. Weather forecasts are a special type of texts which use a very specific language style. This work is usually carried out by hand by meteorologists, which is expensive, time-consuming, and subjected to human errors.

We have introduced a new approach where an ontology, which models the meteorological knowledge, is in charge of guiding the automatic data extraction from the Web to verify them. This ontology includes information about the extraction methods to be applied for each meteorological variable, and also decides how to split the forecasts by the meteorological variables. The architecture lets embed several methods to perform the data extraction. The first tests seem to indicate that using semantics tools to guide the extraction process improves the results obtained by other approaches, and makes them usable in a real environment.

Technical Considerations.

In this context, we deal with a particular type of weather forecasts: those expressed in natural language. We considered the use of machine learning techniques, but in agreement with other studies, we think it is not advised in these kind of contexts because it is the typical case of use of a text belonging to a closed domain.

We have taken advantage of the Semantic Web tools to design an ontology-driven approach which improves the results of the classical approach. The ontology is used as a tool to split the weather forecasts according to the atmospheric variables, in order to use the most appropriate extraction methods. This way of working avoids errors due to ambiguous language.

Moreover, it allows us to integrate any extraction method dynamically with the system, ranging from custom made parsers based on easy rules (e.g., detecting the presence of certain keywords), to more complex ones (e.g., using complex rules or statistical approaches). Besides, this methodology based on placing the extraction parameters in an external ontology facilitates the maintenance labors and the evolution of the system because the changes, adjustments, and extensions can be done in an easier way.

Meteorological Aspects.

Regarding meteorological aspects, there are many ways of using the information extracted from the forecasts. These could be some possibilities:

- To verify the level of accuracy. The data can be segmented according to the forecaster, the geographical area, or a specific atmospheric variable, to name a few.
- To study the use of the style guides by forecasters, in order to check ambiguities in the predictions, the frequency of use of some keywords, or the geographical features which they mention. This would allow to fix common problems found in worded forecasts, such as the heterogeneity of the predictions (by different human forecasters), the bad usage of the language, or the too low (or excessive) risk taken by forecasters when predicting a difficult meteorological situation.
- To verify data over time, in order to identify particular trends over different time periods, and to localize unconscious biases which can plague the predictions.

Open tasks.

Our next steps go first to complete in detail the experiments, establishing a baseline with current technology, and checking the improvement experienced with our approach. From a meteorological point of view, it would be interesting improve extraction methods to ensure the reliability of extraction, particularly in forecasts where major meteorological phenomena are involved. For instance, an extraordinary temperature increase, or a very heavy rainfall. With regard to verification, a very deep study of the results should be made to draw conclusions from a purely meteorological perspective, which will be the subject of further works. Therefore, we do not intend with this work to evaluate the accuracy of the work of meteorologists, but offer an integrated solution in order to help to improve the quality of weather forecasts in the future.

ACKNOWLEDGEMENTS

This research work has been supported by the CICYT project TIN2013-46238-C4-4-R, and DGA-FSE. The authors also wish to thank Dr. Eduardo Mena and AEMET for their support.

REFERENCES

Aquin, M. D., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., and Guidi, D. (2008). Toward a new generation of semantic web applications. *Intelligent Systems, IEEE*, 23(3):20–28.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T. (2011). Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. *Monthly Weather Review*, 139(12):3887–3905.

Belz, A. (2008). Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(04):431–455.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

Bobed, C., Yus, R., Bobillo, F., Ilarri, S., Bernad, J., Mena, E., Trillo-Lado, R., and Garrido, Á. L. (2016). Emerging semantic-based applications. In *Semantic Web*, pages 39–83. Springer.

Borobia, J. R., Bobed, C., Garrido, A. L., and Mena, E. (2014). SIWAM: Using social data to semantically assess the difficulties in mountain activities. In *10th International Conference on Web Information Systems and Technologies (WEBIST'14)*, pages 41–48.

Breslin, J. G., O'Sullivan, D., Passant, A., and Vasiliu, L. (2010). Semantic web computing in industry. *Computers in Industry*, 61(8):729–741.

Buey, M. G., Garrido, Á. L., and Ilarri, S. (2014). An approach for automatic query expansion based on nlp and semantics. In *Database and Expert Systems Applications*, pages 349–356. Springer.

Done, J., Davis, C. A., and Weisman, M. (2004). The next generation of nwp: Explicit forecasts of convection using the weather research and forecasting (wrf) model. *Atmospheric Science Letters*, 5(6):110–117.

Garrido, A., Gómez, O., Ilarri, S., and Mena, E. (2011). NASS: News Annotation Semantic System. In *23rd International Conference on Tools with Artificial Intelligence (ICTAI 2011)*, pages 904–905. IEEE.

Goldberg, E., Driedger, N., and Kittredge, R. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

He, Y., Wetterhall, F., Cloke, H., Pappenberger, F., Wilson, M., Freer, J., and McGregor, G. (2009). Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteorological Applications*, 16(1):91–101.

Horrocks, I. (2008). Ontologies and the semantic web. *Communications of the ACM*, 51(12):58–67.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, 30(4):291–303.

Mathiesen, P. and Kleissl, J. (2011). Evaluation of numerical weather prediction for intra-day solar forecasting in the continental united states. *Solar Energy*, 85(5):967–977.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.

Murphy, A. H. and Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7(4):435–455.

Reiter, E., Sripada, S., Hunter, J., Yu, J., and Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169.