INLG 2022

Centro Singular de Investigación
en **Tecnoloxías Intelixentes**

# Dealing with hallucination and omission in neural NLG: A use case on meteorology

**Javier González-Corbelle, Jose M. Alonso-Moral, A. Bugarín-Diz**
Centro Singular en Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
*{j.gonzalez.corbelle, josemaria.alonso.moral, alberto.bugarin.diz}@usc.es*

**J. Taboada**
MeteoGalicia, Xunta de Galicia,
Santiago de Compostela, Spain
*coordinador-predicion.meteogalicia@xunta.gal*

# Table of contents

- **Motivation & Context**
- **Contributions**
  - MeteoGalicia-ES Dataset
  - Data-to-text generator
  - Divergences analysis
- **Conclusion & Future work**

# Motivation

## Recycling and Reusing Large Language Models (GPT-3, BERT,...)

## Can Experts Trust on Neural NLG models?

# Context

- **Meteorology field**
  - Previous collaborations with MeteoGalicia: fully operational template-based generator
  - Data and expert availability for assessing the quality of the generated texts

**Would a meteorologist use a neural NLG system in a production environment?**

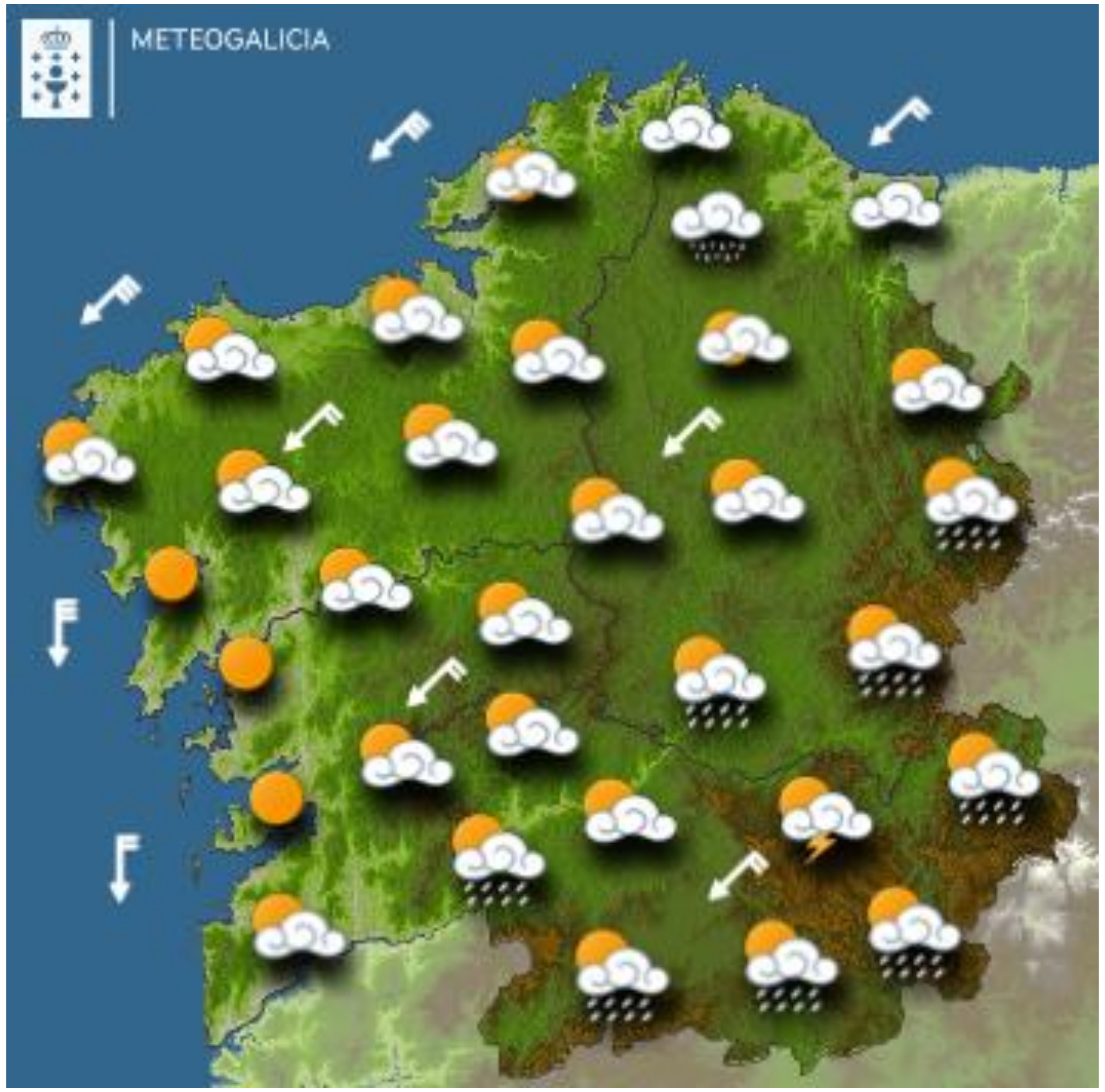
www.meteogalicia.gal

# Table of contents

- Motivation & Context
- **Contributions**
  - **MeteoGalicia-ES Dataset**
  - Data-to-text generator
  - Divergences analysis
- Conclusion & Future work

# MeteoGalicia-ES Dataset

- Data-to-text dataset
- 3033 records of meteorological tabular data
- Real state-of-the-sky data from the community of Galicia (from 2010 to 2020)
- Handwritten textual descriptions in Spanish



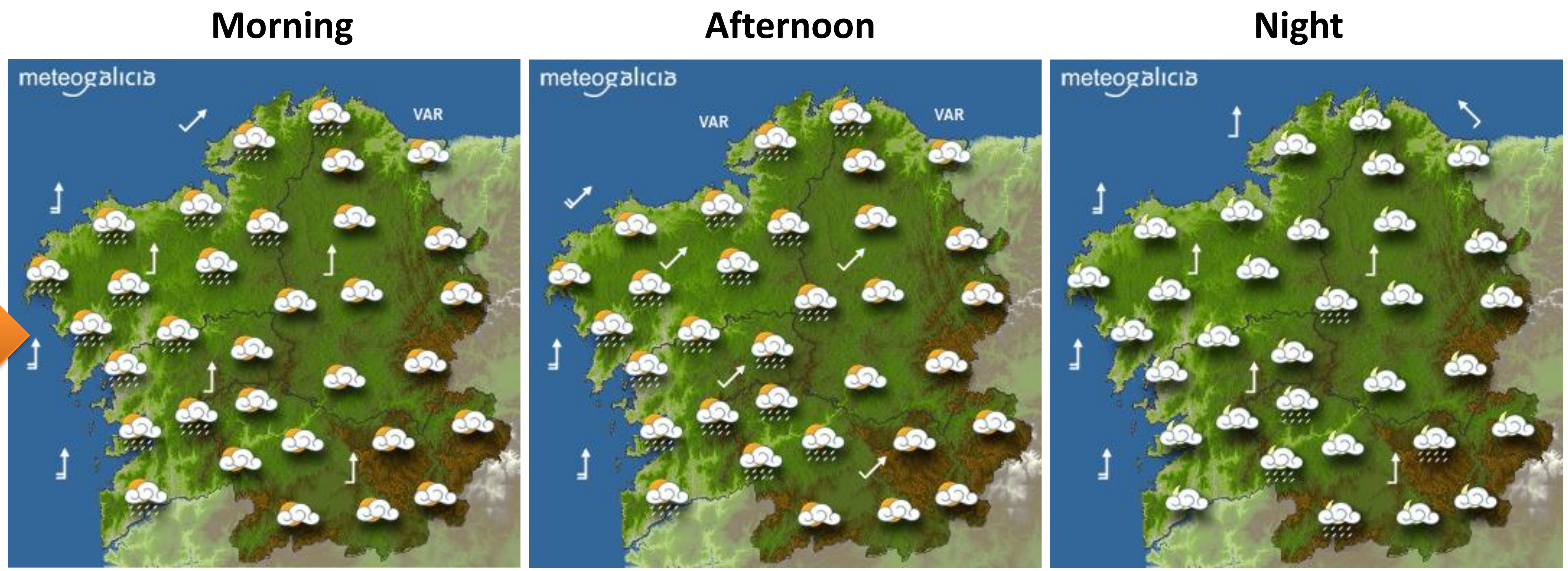
- Publicly available at https://gitlab.citius.usc.es/gsi-nlg/meteogalicia-es

# MeteoGalicia-ES Dataset

- 96 state-of-the-sky values per table

32

| Zone | Morning | Afternoon | Night |
|---|---|---|---|
| A Coruña | Cloudy with showers | Cloudy with showers | Cloudy with sunny spells |
| Pontevedra | Cloudy with showers | Cloudy with showers | Covered |
| … | … | … | … |
| Deza | Cloudy with sunny spells | Cloudy with showers | Cloudy with sunny spells |



Morning

Afternoon

Night

**Reference text:** Clouds and some weak rains. Thus, the skies will be cloudy throughout the Community, with intermittent showers more frequent in the provinces of A Coruña and Pontevedra.

# MeteoGalicia-ES Dataset

- Data tables: **State-of-the-sky categorical values**

| | | | | | |
|---|---|---|---|---|---|
| Clear (47.25%) | | Clear | | Sunny intervals | Cloud (93.74%) |
| Rain (68.84%) | | Weak rains | | Clouds | |
| | | Rains | | High clouds | |
| | | Weak showers | | Cloudy with sunny spells | |
| | | Showers | | Covered | |
| | | Drizzle | | Snow showers | Snow (14.41%) |
| | | Cloudy with showers | | Snow | |
| Fog (40.45%) | | Fog | | Hail | |
| | | Fog banks | | Sleet | |
| | | Mist | | Storm | Storm (8.41%) |

# MeteoGalicia-ES Dataset

Dataset statistics: texts

- Value references: values from the table (e.g., “clear”, “clouds”)
- Spatial references: following MeteoGalicia’s official style guide (e.g., “north coast”, “west”)
- Temporal references: preliminary search of words “morning”, “afternoon” and “night”

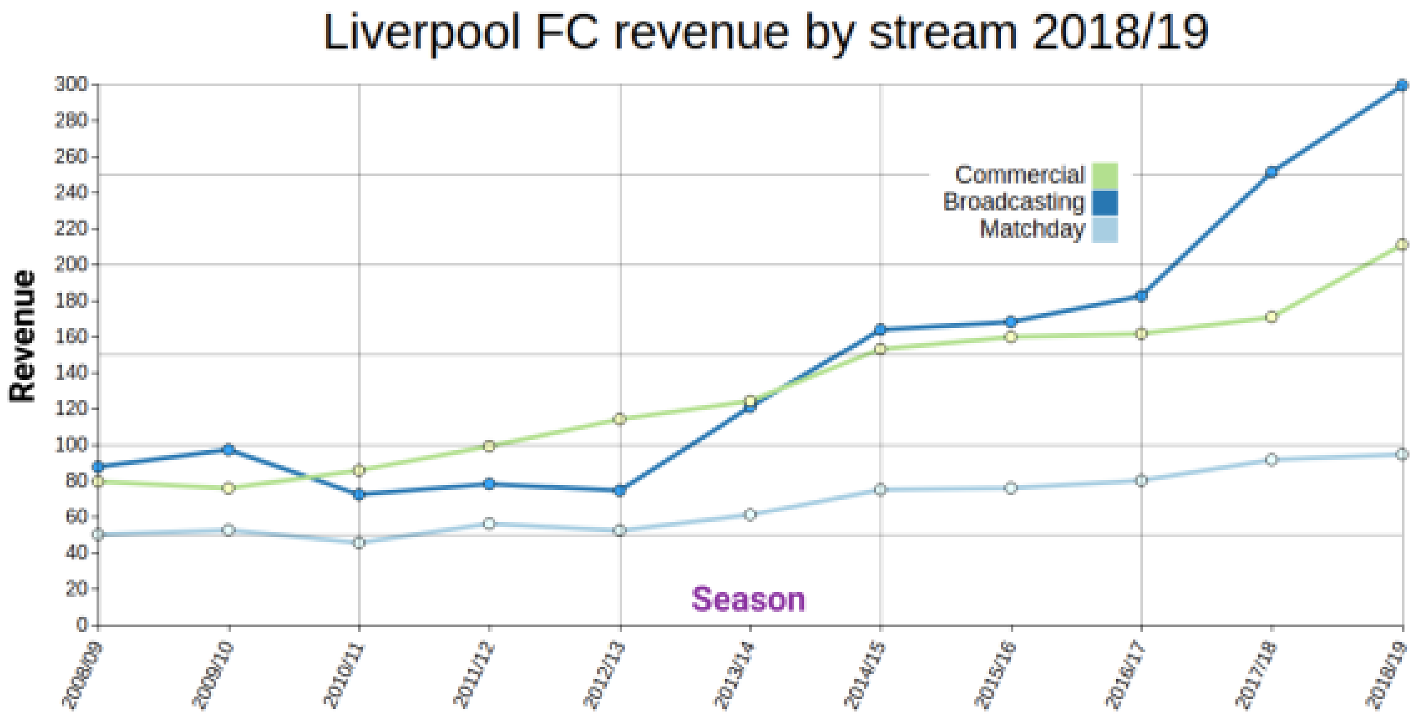
| References per text | | |
|---|---|---|
| **Value refs.** | **Spatial refs.** | **Temporal refs.** |
| 2.53 | 1.5 | 1.07 |

# Table of contents

- Motivation & Context
- **Contributions**
  - MeteoGalicia-ES Dataset
  - **Data-to-text generator**
  - Divergences analysis
- Conclusion & Future work

# Data-to-text generator

- Base model: **Chart-to-text** (Transformer-based model)



The broadcasting revenue stream is the largest source of revenue for Liverpool FC. In 2018/2019, the football club earned approximately 299.3 million euros from domestic and international competitions broadcasting, more than twice of what they earned in 2011/2012. The second biggest revenue stream was commercial – sponsorships and merchandising.

## Model Output:

The templateLabel[2][0] templateTitle[2] templateTitle[4] is the largest source of templateTitle[2] for templateTitleSubject[0]. In 2018/2019, the football club earned approximately templateValue[2][0] templateScale euros from domestic and international competitions templateLabel[2][0], more than twice of what they earned in 2011/2012. The second biggest templateTitle[2] templateTitle[4] was templateLabel[3][0] – sponsorships and merchandising.

## After Variable Substitution:

The Broadcasting revenue stream is the largest source of revenue for Liverpool FC . In 2018/2019 , the football club earned approximately 299.3 million euros from domestic and international competitions Broadcasting , more than twice of what they earned in 2011/2012 . The second biggest revenue stream was Commercial – sponsorships and merchandising .

J. Obeid, E. Hoque, "**Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model**", INLG2020. https://aclanthology.org/2020.inlg-1.20/

# Data-to-text generator

## Adapting chart-to-text to MeteoGalicia-ES

- Input data and pre-processing
  - Chart title ↔ Generic title (“Weather forecast of a day in Galicia, by period of the day”)
  - NER in Spanish
- Training and validation: 75-15-15 partition, maintain same layers and parameters
- Testing and post-processing
  - Model generates templates, and we add BETO to fill in the gaps in Spanish
  - Post-processing corrections (mostly typos)

J. Cañete, et. al., “Spanish pre-trained BERT model and evaluation data”, PML4DC at ICLR 2020.

# Data-to-text generator

## Generation

- Input data: 3,033 records from MeteoGalicia-ES
- Pre-processing step: before training, the model cleans our dataset and lefts 1,815 samples
  - Training samples: 1,270
  - Validation samples: 273
  - Test samples: 272 to analyze and evaluate

# Table of contents

- Motivation & Context
- **Contributions**
  - MeteoGalicia-ES Dataset
  - Data-to-text generator
  - **Divergences analysis**
- Conclusion & Future work

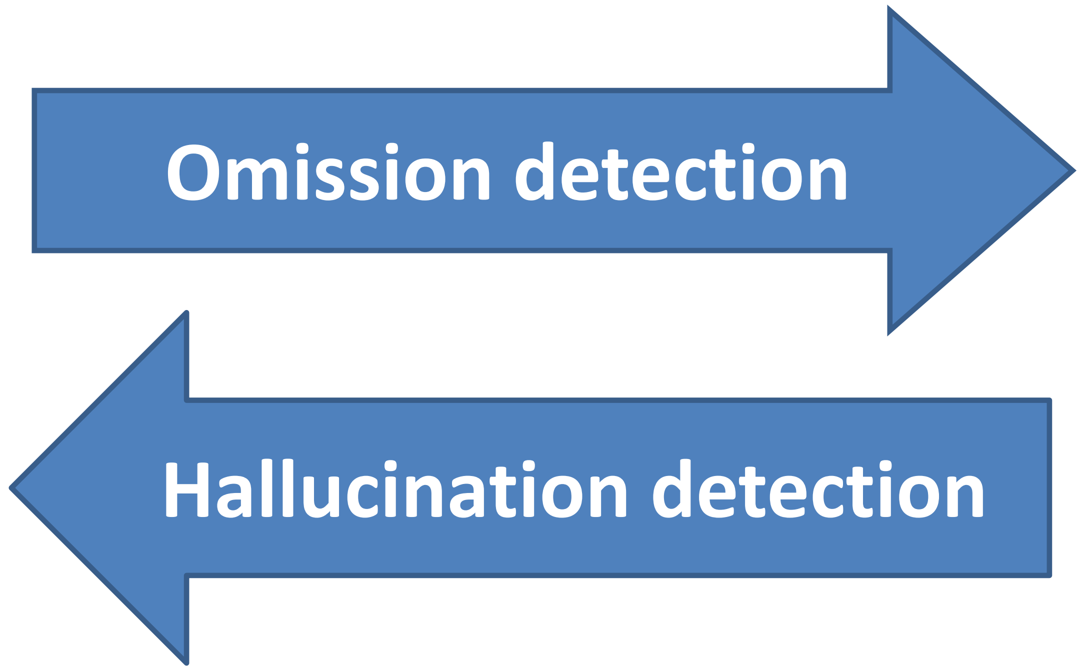
# Divergences analysis

## Divergence detector

- **Omissions**: input information not mentioned in the generated caption
- **Hallucinations**: caption mentions information not in the data input

| Zone | Morning | Afternoon | Night |
|---|---|---|---|
| Mariña Oriental | Cloudy with sunny spells | Cloudy with sunny spells | Cloudy with sunny spells |
| Mariña Occidental | Cloudy with showers | Cloudy with showers | Cloudy with sunny spells |
| … | … | … | … |
| Deza | Cloudy with sunny spells | Cloudy with showers | Cloudy with sunny spells |

Are all the input values mentioned in the generated text?



**Generated text:** The skies will be cloudy in general, with intermittent showers, more frequent in the coast.

1. Are all the mentioned values in the input data?
2. Are all the mentioned values referring the correct location? (Spatial hallucination)

# Divergences analysis

## Detected divergences

- Omissions: 160 out of 272 (58%)
- Hallucinations
  - Basic: 35 out of 272 (12.9%)
  - Spatial: 11 out of 272 (4%)
- **How many of these divergences are admissible in our context?**

Dušek et al., 2019, “Semantic Noise Matters for Neural Natural Language Generation”, INLG 2019.

# Divergences analysis

**Expert evaluation of the detected hallucinations**

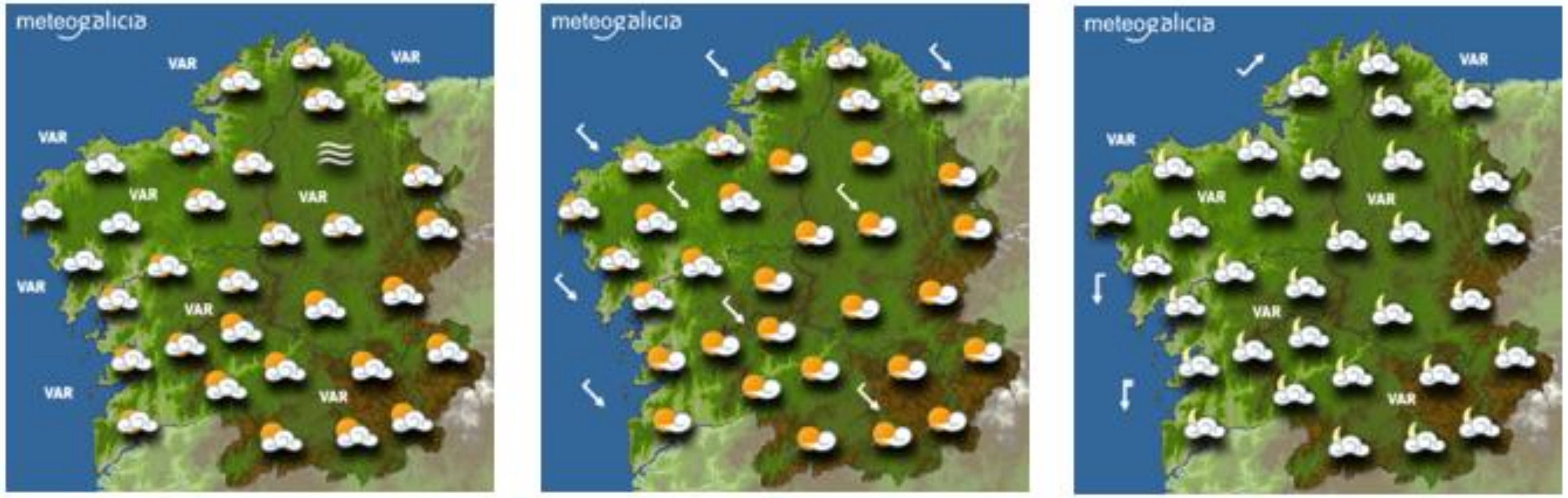
- Rate degree of relevance out of a 3-point Likert scale
  - Admissible: the text is consistent with the meteorological situation and can be considered as correct, despite not being perfect.
  - Partly admissible: the text is not entirely faithful to the data; it could refer to the state-of-the-sky values and the associated area more accurately.
  - Inadmissible: the text mentions values not in the data or associated to a wrong area.
- Comment: justify the score

# Divergences analysis

Generated text: Skies will be very cloudy in the morning with cloudy and clear skies and some weak rains. In the afternoon, very open skies will prevail.

Hallucinated values: {'weak rains'}

Data values (input): {'cloudy with sunny spells', 'covered', 'clouds', 'high clouds', 'fog'}



| | | | |
|---|---|---|---|
| **Score** | ADMISSIBLE<br>☒ | PARTLY ADMISSIBLE<br>□ | INADMISSIBLE<br>□ |
| **Comment (optional)** | No showers should be expected, but with areas of complete overcast it is not out of place to mention the possibility of some light rain. | | |

# Divergences analysis

## Results

| | Admissible | Partly Admissible | Inadmissible | Total |
|---|---|---|---|---|
| **Basic hallucinations** | 10 (28.6%) | 13 (37.1%) | 12 (34.3%) | 35 |
| **Spatial hallucinations** | 2 (18.2%) | 4 (36.3%) | 5 (45.5%) | 11 |
| **Total hallucinations** | 12 (26%) | 17 (37%) | 17 (37%) | 46 |

**Contextual information + Meteorologist's background + Commonsense reasoning**

# Conclusions

- Reusing a data-to-text neural model is not a trivial task
- Lack of **domain knowledge** produces unfaithful content
- Depending on the context, hallucinations could be admissible

## Can Experts Trust on Neural NLG models?

- Meteorologists don’t trust yet, they prefer a less natural text, but reliable

# Future work

Improve both the data-to-text system and the detector:

- **Mid-term**: enrich with a meteorological knowledge base in agreement with the experts
  - Spatial references
  - Temporal references (at day level)
- **Long-term**: include temporal knowledge (historical knowledge)

INLG 2022

Centro Singular de Investigación en Tecnoloxías Intelixentes

# Dealing with hallucination and omission in neural NLG: A use case on meteorology

**Javier González-Corbelle, Jose M. Alonso-Moral, A. Bugarín-Diz**
Centro Singular en Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
*{j.gonzalez.corbelle, josemaria.alonso.moral, alberto.bugarin.diz}@usc.es*

**J. Taboada**
MeteoGalicia, Xunta de Galicia,
Santiago de Compostela, Spain
*coordinador-predicion.meteogalicia@xunta.gal*