Meteorologists and Students: A resource for language grounding of geographical descriptors

Alejandro Ramos-Soto[1,2], Ehud Reiter[2], Kees van Deemter[2,3], Jose M. Alonso[1], and Albert Gatt[4]

[1]Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela:
{alejandro.ramos,josemaria.alonso.moral}@usc.es
[2]Department of Computing Science, University of Aberdeen:
{alejandro.soto,e.reiter,k.vdeemter}@abdn.ac.uk
[3]Utrecht University: k.vandeemter@uu.nl
[4]Institute of Linguistics and Language Technology, University of Malta: albert.gatt@um.edu.mt

Abstract

We present a data resource which can be useful for research purposes on language grounding tasks in the context of geographical referring expression generation. The resource is composed of two data sets that encompass 25 different geographical descriptors and a set of associated graphical representations, drawn as polygons on a map by two groups of human subjects: teenage students and expert meteorologists.

1 Introduction

Language grounding, i.e., understanding how words and expressions are anchored in data, is one of the initial tasks that are essential for the conception of a data-to-text (D2T) system (Roy and Reiter, 2005; Reiter, 2007). This can be achieved through different means, such as using heuristics or machine learning algorithms on an available parallel corpora of text and data (Novikova et al., 2017) to obtain a mapping between the expressions of interest and the underlying data (Reiter et al., 2005), getting experts to provide these mappings, or running surveys on writers or readers that provide enough data for the application of mapping algorithms (Ramos-Soto et al., 2017).

Performing language grounding ensures that generated texts include words whose meaning is aligned with what writers understand or what readers would expect (Roy and Reiter, 2005), given the variation that is known to exist among writers and readers (Reiter and Sripada, 2002). Moreover, when contradictory data appears in corpora or any other resource that is used to create the data-to-words mapping, creating models that remove inconsistencies can also be a challenging part of language grounding which can influence the development of a successful system (Reiter et al., 2005).

This paper presents a resource for language grounding of geographical descriptors. The original purpose of this data collection is the creation of models of geographical descriptors whose meaning is modeled as graded or fuzzy (Fisher, 2000; Fisher et al., 2006), to be used for research on generation of geographical referring expressions, e.g., (Turner et al., 2010, 2008; de Oliveira et al., 2015; Ramos-Soto et al., 2016, 2017). However, we believe it can be useful for other related research purposes as well.

2 The resource and its interest

The resource is composed of data from two different surveys. In both surveys subjects were asked to draw on a map (displayed under a Mercator projection) a polygon representing a given geographical descriptor, in the context of the geography of Galicia in Northwestern Spain (see Fig. 1). However, the surveys were run with different purposes, and the subject groups that participated in each survey and the list of descriptors provided were accordingly different.

The first survey was run in order to obtain a high number of responses to be used as an evaluation testbed for modeling algorithms. It was answered by 15/16 year old students in a high school in Pontevedra (located in Western Galicia). 99 students provided answers for a list of 7 descriptors (including cardinal points, coast, inland, and a proper name). Figure 2 shows a representation of the answers given by the students for "Northern Galicia" and a contour map that illustrates the percentages of overlapping answers.

The second survey was addressed to meteorologists in the Galician Weather Agency (MeteoGalicia, 2018). Its purpose was to gather data to create

Proceedings of The 11th International Natural Language Generation Conference, pages 421–425,
Tilburg, The Netherlands, November 5-8, 2018.

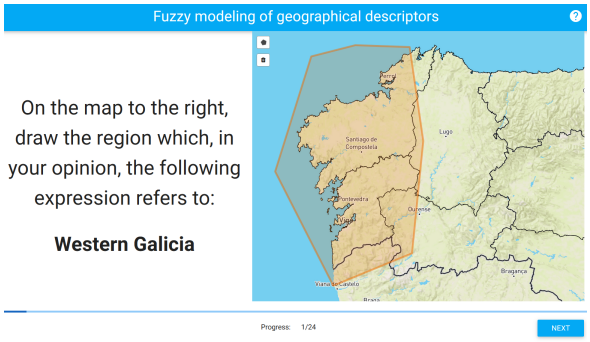

Figure 1: Snapshot of the version of the survey answered by the meteorologists (translated from Spanish).

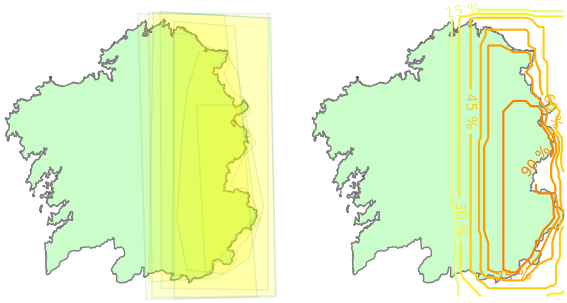

Figure 3: Representation of polygon drawings by experts and associated contour plot showing the percentage of overlapping answers for "Eastern Galicia".

fuzzy models that will be used in a future NLG system in the weather domain. Eight meteorologists completed the survey, which included a list of 24 descriptors. For instance, Figure 3 shows a representation of the answers given by the meteorologists for "Eastern Galicia" and a contour map that illustrates the percentage of overlapping answers.

Table 1 includes the complete list of descriptors for both groups of subjects. 20 out of the 24 descriptors are commonly used in the writing of weather forecasts by experts and include cardinal directions, proper names, and other kinds of references such as mountainous areas, parts of provinces, etc. The remaining four were added to study intersecting combinations of cardinal directions (e.g. exploring ways of combining "north" and "west" for obtaining a model that is similar to "northwest").

The data for the descriptors from the surveys is focused on a very specific geographical context. However, the conjunction of both data sets provides a very interesting resource for performing a variety of more general language grounding-oriented and natural language generation research tasks, such as:

- Testing algorithms that create geographical models. These models would aggregate the answers from different subjects for each descriptor. The differences among the subjects can be interpreted from a probabilistic or fuzzy perspective that allows a richer characterization of the resulting models. For instance, in Fig. 2 the contour plots could be taken as the basis or support for the semantics of the expression "Northern Galicia", with a core region that is accepted by the majority, and a gradual decay as one moves to the outer periphery of the regions outlined.
- Analyzing differences between the expert and non-expert groups for the descriptors they have in common (as Table 1 shows, both groups share 6 descriptors).
- Studying how to combine models represent-

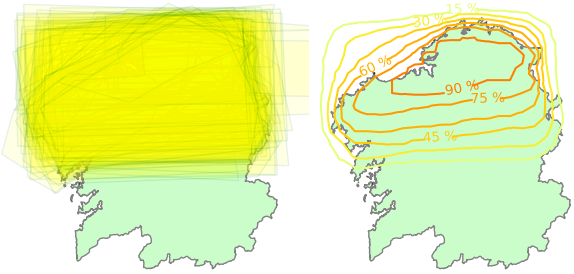

Figure 2: Representation of polygon drawings by students and associated contour plot showing the percentage of overlapping answers for "Northern Galicia".

Subject Group	Spanish	English translation
Common	Norte de Galicia, Sur de Galicia, Oeste de Galicia, Este de Galicia, Interior de Galicia, Rías Baixas	Northern Galicia, Southern Galicia, Western Galicia, Eastern Galicia, Inland Galicia, Rías Baixas
Students	Costa de Galicia	Galician Coast
Experts	Tercio Norte, Extremo Norte Áreas de montaña de Lugo, Áreas de montaña de Ourense, Oeste de A Coruña, Comarcas Atlánticas, Litoral Atlántico, Litoral Cantábrico, Litoral Norte, Interior de Coruña, Interior de Pontevedra, Oeste de Ourense, Sur de Ourense, Sur de Lugo, Noroeste de Galicia, Noreste de Galicia, Suroeste de Galicia, Sureste de Galicia	Northern Third, Extreme North, Mountainous areas in Lugo, Montainous areas in Ourense, Western A Coruña, Atlantic Regions, Atlantic Coast, Cantabrian Coast, Northern Coast, Inland Coruña, Inland Pontevedra, Western Ourense, Southern Ourense, Southern Lugo, Northwestern Galicia, Northeastern Galicia, Southwestern Galicia, Southeastern Galicia

Table 1: List of geographical descriptors in the resource.

ing the semantics of different cardinal directions, such as “south” and “east” to obtain a representation of “southeast”.

- Developing geographical referring expression generation algorithms based on the empirically created models.

3 Qualitative analysis of the data sets

The two data sets were gathered for different purposes and only coincide in a few descriptors, so providing a direct comparison is not feasible. However, we can discuss general qualitative insights and a more detailed analysis of the descriptors that both surveys share in common.

At a general level, we had hypothesized that experts would be much more consistent than students, given their professional training and the reduced number of meteorologists participating in the survey. Comparing the visualizations of both data sets we have observed that this is clearly the case; the polygons drawn by the experts are more concentrated and therefore there is a higher agreement among them. On top of these differences, some students provided unexpected drawings in terms of shape, size, or location of the polygon for several descriptors.

If we focus on single descriptors, one interesting outcome is that some of the answers for “Northern Galicia” and “Southern Galicia” overlap for both subject groups. Thus, although ‘north’ and ‘south’ are natural antonyms, if we take into account the opinion of each group as a whole, there exists a small area where points can be considered as belonging to both descriptors at the same time (see Fig. 4). In the case of “west” and “east”, the drawings made by the experts were almost divergent and showed no overlapping between those two descriptors.

Regarding “Inland Galicia”, the unions of the answers for each group occupy approximately the same area with a similar shape, but there is a very high overlapping among the answers of the meteorologists. A similar situation is found for the remaining descriptor “Rías Baixas”, where both groups encompass a similar area. In this case, the students’ answers cover a more extensive region and the experts coincide within a more restricted area.

3.1 A further analysis: apparent issues

As in any survey that involves a task-based collection of data, some of the answers provided by the subjects for the described data sets can be considered erroneous or misleading due to several reasons. Here we describe for each subject group some of the most relevant issues that any user of this resource should take into account.

In the case of the students, we have identified minor drawing errors appearing in most of the descriptors, which in general shouldn’t have a negative impact in the long term thanks to the high number of participants in the original survey. For some descriptors, however, there exist polygons drawn by subjects that clearly deviate from what could be considered a proper answer. The clearest example of this problem involves the ‘west’ and ‘east’ descriptors, which were confused by some of the students who drew them inversely (see Fig. 5, around 10-15% of the answers).

In our case, given their background, some of the students may have actually confused the meaning of + “west” and “east”. However, the most plausible explanation is that, unlike in English and other languages, in Spanish both descriptors are phonetically similar (“este” and “oeste”) and can be easily mistaken for one another if read without atten-

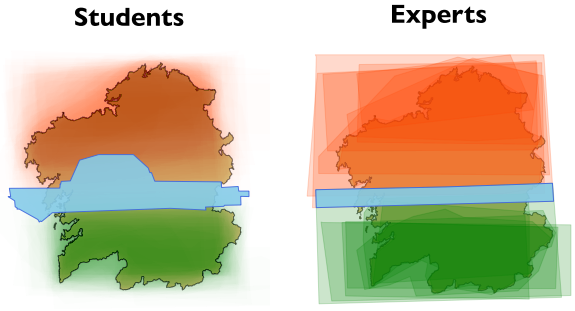

Figure 4: Areas overlapping “north” and “south” for both subject groups (in blue).

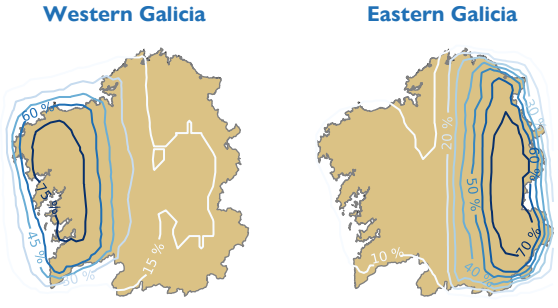

Figure 5: Contour maps of student answers for “Western Galicia” and “Eastern Galicia”.

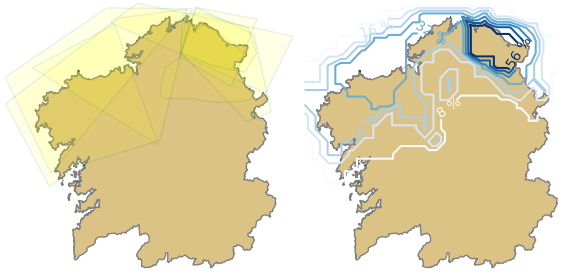

Figure 6: Representation of polygon drawings by experts and associated contour plots showing the percentage of overlapping answers for "Northeastern Galicia".

tion.

As for the expert group, a similar case is found for "Northeastern Galicia" (see Fig. 6), where some of the given answers (3/8) clearly correspond to "Northwestern Galicia". However, unlike the issue related to "west" and "east" found for the student group, this problem is not found reciprocally for the "northwestern" answers.

4 Resource materials

The resource is available at (Ramos-Soto et al., 2018) under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Both data sets are provided as SQLite databases which share the same table structure, and also in a compact JSON format. Polygon data is encoded in GeoJSON format (Butler et al., 2016). The data sets are well-documented in the repository's README, and several Python scripts are provided for data loading, using Shapely (Gillies et al., 2007–2018); and for visualization purposes, using Cartopy (Met Office, 2010–2015).

5 Concluding remarks

The data sets presented provide a means to perform different research tasks that can be useful from a natural language generation point of view. Among them, we can highlight the creation of models of geographical descriptors, comparing models between both subject groups, studying combinations of models of cardinal directions, and researching on geographical referring expression generation. Furthermore, insights about the semantics of geographical concepts could be inferred under a more thorough analysis.

One of the inconveniences that our data sets present is the appearance of the issues described in Sec. 3.1. It could be necessary to filter some of the answers according to different criteria (e.g., deviation of the centroid location, deviation of size, etc.). For more applied cases, manually filtering can also be an option, but this would require a certain knowledge of the geography of Galicia. In any case, the squared-like shape of this region may allow researchers to become rapidly familiar with many of the descriptors listed in Table 1.

As future work, we believe it would be invaluable to perform similar data gathering tasks for other regions from different parts of the world. These should provide a variety of different shapes (both regular and irregular), so that it can be feasible to generalize (e.g., through data-driven approaches) the semantics of some of the more common descriptors, such as cardinal points, coastal areas, etc. The proposal of a shared task could help achieve this objective.

Acknowledgments

This research was supported by the Spanish Ministry of Economy and Competitiveness (grants TIN2014-56633-C3-1-R and TIN2017-84796-C2-1-R) and the Galician Ministry of Education (grants GRC2014/030 and "accreditation 2016-2019, ED431G/08"). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program). A. Ramos-Soto is funded by the "Consellería de Cultura, Educación e Ordenación Universitaria" (under the Postdoctoral Fellowship accreditation ED481B 2017/030). J.M. Alonso is supported by RYC-2016-19802 (Ramón y Cajal contract).

The authors would also like to thank Juan Taboada for providing the list of most frequently used geographical expressions by MeteoGalicia, and José Manuel Ramos for organizing the survey at the high school IES Xunqueira I in Pontevedra, Spain.

References

H. Butler, M. Daly, A. Doyle, S. Gillies, S. Hagen, and T. Schaub. 2016. *"The GeoJSON Format", RFC 7946*.

Peter Fisher. 2000. Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, 113(1):7–18.

Peter Fisher, Alexis Comber, and Richard Wadsworth. 2006. Approaches to uncertainty in spatial data. *Fundamentals of Spatial Data Quality*, pages 43–59.

Sean Gillies et al. 2007–2018. Shapely: manipulation and analysis of geometric objects. https://github.com/Toblerity/Shapely.

Met Office. 2010–2015. Cartopy: a cartographic python library with a matplotlib interface. http://scitools.org.uk/cartopy.

MeteoGalicia. 2018. Meteogalicia's web site. http://www.meteogalicia.es.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Saarbrücken, Germany. ArXiv:1706.09254.

Rodrigo de Oliveira, Yaji Sripada, and Ehud Reiter. 2015. Designing an algorithm for generating named spatial references. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 127–135. Association for Computational Linguistics.

Alejandro Ramos-Soto, Jose M Alonso, Ehud Reiter, Kees van Deemter, and Albert Gatt. 2017. An empirical approach for modeling fuzzy geographical descriptors. In *Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on*, pages 1–6. IEEE.

Alejandro Ramos-Soto, Ehud Reiter, Kees van Deemter, Jose M. Alonso, and Albert Gatt. 2018. geodescriptors. https://gitlab.citius.usc.es/alejandro.ramos/geodescriptors. Accessed: 2018-07-09.

Alejandro Ramos-Soto, Nava Tintarev, Rodrigo de Oliveira, Ehud Reiter, and Kees van Deemter. 2016. Natural language generation and fuzzy sets: An exploratory study on geographical referring expression generation. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 587–594.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the 11th European Workshop on Natural Language Generation*, pages 97–104.

Ehud Reiter and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of the International Natural Language Generation Conference*, pages 97–104.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Ross Turner, Somayajulu Sripada, and Ehud Reiter. 2010. Generating approximate geographic descriptions. In *Empirical Methods in Natural Language Generation*, volume 5790 of *Lecture Notes in Computer Science*, pages 121–140. Springer Berlin Heidelberg.

Ross Turner, Somayajulu Sripada, Ehud Reiter, and Ian P. Davy Davy. 2008. Using spatial reference frames to generate grounded textual summaries of georeferenced data. In *Proceedings of the 2008 International Conference on Natural Language Generation (INLG08)*, pages 16–24.