Absolute and Relative Properties in Geographic Referring Expressions

Rodrigo de Oliveira, Somayajulu Sripada and **Ehud Reiter**
University of Aberdeen
{rodrigodeoliveira, yaji.sripada, e.reiter}@abdn.ac.uk

Abstract

This paper discusses the importance of computing relative properties and not just retrieving absolute properties when generating geographic referring expressions such as "northern France". We describe an algorithm that computes spatial properties at run-time by means of spatial operations such as intersecting and analyzing parts of wholes. The evaluation of the algorithm suggests that part-whole relations are key in geographic expressions.

1 Introduction

This paper discusses the role of spatial operations in 'creating' properties to be used for generating geographic expressions. For example, we generate the expression "northern France" by retrieving the property FRANCE from our knowledge base, and subsequently computing (or creating) the property NORTH at run-time. The algorithm we describe in this article is meant to be used by Natural Language Generation (NLG) systems (Reiter and Dale, 2000), especially those in the Data-to-Text family (Reiter, 2007), which automatically write reports in natural language such as English, given structured data such as those we typically store in databases. Our domain is weather forecast and our input data conforms with that typically found in Geographic Information Systems (Worboys and Duckham, 2004).

The many algorithms for doing Referring Expression Generation (REG) as outlined in Krahmer and Van Deemter (2012) assume that Knowledge Bases (KBs) exhaustively specify all properties that are inherent (i.e. absolute) to entities. The REG style we propose here is inspired in alternative work (Kelleher and Kruijff, 2006; Viethen and Dale, 2008) that computes relational properties, rather than storing them in KBs. We base our approach on evidence observed in human-authored texts, as it shall be explained in Section 4. The underlying philosophy is that some properties are absolute, i.e. inherent to entities, while some properties are relative to other properties. An example of the relative type of properties in the spatial domain is the part-whole relation, henceforth *mereology* (Cohn and Renz, 2008, 577). For example, a given city will absolutely be a part of a country (or continent) or not, so the properties COUNTRY and CONTINENT are absolute. On the other hand, whether a city lies in the North depends on the area that is chosen as the whole, so the property DIRECTION is relative to another property. Paris is in the North of France, but lies in the centre of Europe. NORTH and CENTRAL are in a mereological relation to FRANCE and EUROPE, respectively.

Our approach is very much in line with that proposed by Van Deemter (2002), since we process sets (not individuals) by computing intersection, a typical set-theoretic operation. The key difference from a fully set-theoretic approach is that we also compute mereological relations. As described in Sections 2 and 3, our algorithm takes point-based data and outputs sets of semantic labels such as (COASTAL $\sqcap$ (NORTH, FRANCE)). Such sets can be further converted into a natural language expression such as "northern coast of France" or "coast in northern France" in a full NLG system. The performance of our approach is evaluated and discussed in Section 5.

Proceedings of The 9th International Natural Language Generation conference, pages 256–264,
Edinburgh, UK, September 5-8 2016. ©2016 Association for Computational Linguistics

2 Concepts Underlying the Algorithm

Before explaining the procedure the algorithm follows, we first need to look at some background concepts that were implemented in the algorithm.

Descriptors are qualitative labels such as NORTH, ABERDEEN, HIGH or COASTAL. When constructing objects representing descriptions, we transform primitive values from the dataset (e.g. elevation=800m) into descriptor labels (e.g. HIGH).

Frames of Reference assign descriptors to particular subsets of the data. Frames are relations between data points and some other spatial entity, using some measurement. Our model ended up with two types of frames, depending on how much the number of relative spatial entities varied:

- **Absolute Frames** are those whose relative spatial entities are few or only one. For instance, whether a point lies on high or low ground always depends (in our domain) on the spatial entity called 'sea' and some arbitrary metric, such as the distance on the z-axis to that entity. This allows descriptors to be labelled as HIGH or LOW, by simply retrieving absolute values of data points. For example, if all points in a subset of points have values above 200 for the property *height*, a descriptor with label HIGH is created to describe that subset. To mimic expressions in our corpus, 3 absolute frames were implemented: COASTALPROXIMITY $\equiv$ (COASTAL $\wedge$ INLAND), ELEVATION $\equiv$ (HIGH $\wedge$ LOW) and NAMEDAREAS $\equiv$ (ABERDEEN $\wedge$ ABERDEENSHIRE $\wedge$ MORAY).

- **Relative Frames** are those whose relative spatial entities are too many, which makes it inappropriate to list all possible relations as potential descriptors of that frame. For example, the 3 regions of NAMEDAREAS (see above) can still be split into compass directions. Assigning a single direction value such as NORTH to a descriptor is ambiguous, since that will depend on the area used as reference. Because the direction of a point in our corpus depends on different spatial entities, we modelled DIRECTIONS as the only relative frame, which contains the 4 cardinal directions (e.g. NORTH) and the 4 inter-cardinal directions (e.g. NORTHEAST)[1].

Geocharacterization is the process of mapping points to descriptors. Geocharacterization creates a finite set of Frames of Reference such as COASTALPROXIMITY and ELEVATION.

Descriptions are sets of descriptors such as (NORTH $\sqcap$ COASTAL)[2] that identify a particular subset of the data. A description never contains more than one descriptor of the same Frame of Reference.

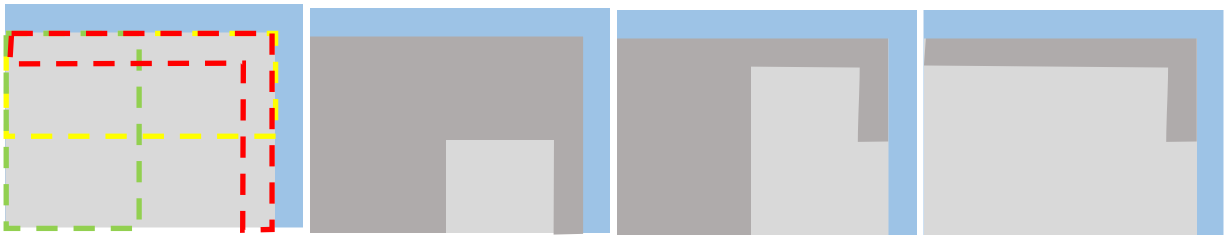
Intersection is the relation between descriptors of a description in which only those points that are common between the descriptors are considered. For example, the description (NORTH $\sqcap$ COASTAL) means that the subset of points being referred to are only those that belong to both NORTH and COASTAL.

Mereology is the relation between descriptors of a description in which a part-whole relation is created, where a named descriptor becomes the whole and a direction descriptor the part. For example, the description (NORTH, ABERDEENSHIRE) implies only the subset of ABERDEENSHIRE we can also label as NORTH. In our approach, we implemented a 4-tile half-panes model (Frank, 1992, 361), where a bounding box is created around a named area. Each half of the box becomes a cardinal direction – the upper half becomes NORTH, the left half WEST, etc., and the intersections between halves become the inter-cardinal directions, e.g. NORTHEAST $\equiv$ NORTH $\sqcap$ EAST.

The concept of Descriptions is particularly important to our approach: they are the representation of geographic referring expressions and are the output of the algorithm. A Description such as (NORTH, COASTAL) can be used by a realiser in an NLG system to generate surface expressions such as "north-

[1] Our dichotomy *absolute vs. relative* does not align with Levinson's relative and absolute frames. We implement frames as functions and call *absolute* those functions that take only the data point as argument (e.g. coastal-proximity(oxford) = inland), and we call *relative* those that take a second argument (e.g. directions(oxford, uk) = south, but directions(oxford, europe) = northwest).

[2] For the sake of readability, when a direction is relative to the entire region, we omit the relation. The description ((NORTH, WHOLE_REGION) $\sqcap$ COASTAL) is simplified to (NORTH $\sqcap$ COASTAL).



(a) Yellow = NORTH, green = EAST, red = COASTAL (b) the North, the West and the coast (c) the northern coast and the West (d) the northern and western coast

Figure 1: Some interpretations of a description (NORTH ? EAST ? COASTAL). To generate expression 1c, our approach needs to output 2 descriptions and unify them: (NORTH ⊓ COASTAL) ⊔ WEST.

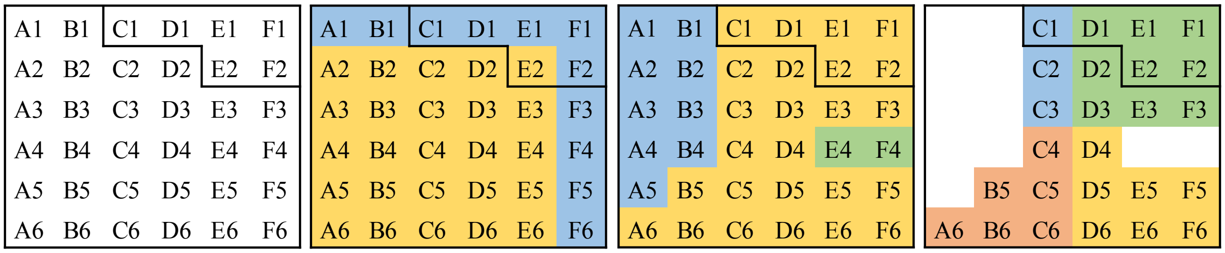
ern coasts", "coasts in the North", "N coast", etc. In our work, we assume such expressions to be surface variations of the same semantic structure. Our algorithm thus outputs a semantic structure (a Description), not a surface form (an expression).

Slightly different forms of the above concepts were used in the work of Turner et al. (2010). However Turner and colleagues limit Frames of Reference (and the set of Descriptors they are made of) to be only absolute, i.e. there is only one specific set of points for each descriptor. Our research, as we explain in more detail below, has shown that this is not true for mereological relations. There is also the danger of selecting content for a referring expression that is not ideal for surface forms as Horacek (2004) and Khan et al. (2008) alert. In the work of Turner and colleagues, descriptions could contain many direction descriptors and the relation between descriptors was not defined (represented as ?). This is harmless for expressions such as "the North and the West", where the description is (NORTH ? EAST). The approach becomes problematic when the final description is (NORTH ? EAST ? COASTAL), as seen in Figure 1. Possible realizations of this description are "the North, the West and the coast", or "the northern coast and the West", "the northern and western coast", among others. Not knowing the relation between the directions and COASTAL enables the system to admit any of these realizations as possible, which could be misleading for a reader. In this paper, we describe mereology as a key spatial relation, but surely others exist. The spatial extension of the Generalized Upper Model (Bateman et al., 2010) lists *internal* and *external* directions, so NORTH could be internal or external to a named area. For example, NORTH is internal in "northern London" (so a mereological relation exists) but it can be either internal or external in "North of London".

It is important to note too that constructing Frames of Reference (i.e. doing geocharacterization) can be influenced by many factors, as suggested by Ramos-Soto et al. (2016), and thus the number of geocharacterization models could be infinite. For instance, the north of regions cannot always be viewed as the absolute upper half of a region. What one calls "North" may depend on many features pertinent to the region. The existence of a mountain range in the middle of an area could become the boundary between north and south. The same applies for coastal proximity. The width of a coastal area may vary depending on the scale with which one looks at a map. We cannot exclude the possibility of geocharacterization variation between individuals either. Therefore we do not claim our specific geocharacterization to be universal; it simply enables us to run an algorithm that should reflect human behaviour when employing spatial operations to generate geographic referring expressions, while leaving geocharacterization models as an open and intriguing question. In other words, our geocharacterization is an *assumption*, and what we carefully investigate is the role of spatial operations in generating geographic expressions.

3 The Algorithm

In this section we explain how our algorithm goes from point-based data to semantic representations of geographic referring expressions. The entire procedure occurs in 2 steps: overgeneration and scoring. The overgeneration step starts with the entire



(a) Raw data. (b) COASTALPROXIMITY. (c) NAMEDAREAS. (d) DIRECTIONS.

Figure 2: Hypothetical geocharacterization models for a region. Model A is the raw data representing the entire region, where the subset {C1, D1, E1, F1, E2, F2} is the target. B represents the COASTALPROXIMITY frame, where blue denotes COASTAL and yellow INLAND. C represents the NAMEDAREAS frame, where blue denotes MORAY, yellow ABERDEENSHIRE and green ABERDEEN. D represents the DIRECTIONS frame for ABERDEENSHIRE. Blue denotes *northwest*, green *northeast*, orange *southwest* and yellow *southeast*. NORTH is the union of *northwest* and *northeast*, EAST the union of *northeast* and *southeast*, and so on.

dataset, which is already tagged with absolute properties (such as named area and altitude). Its goal is to produce all possible descriptions for a subset of the dataset, the target set (e.g. all points where precipitation is observed). At any point, descriptions that do not overlap with the target subset are rejected. The overgeneration algorithm functions as follows:

1. Start a list of candidate descriptions by building single-descriptor descriptions from all absolute frames.

2. Increment the list of candidates with mereological descriptions, i.e. for each NAMEDAREAS descriptor combine it with each relative descriptor (currently only DIRECTIONS descriptors).

3. Increment the list of candidates with all valid intersections[3] among the current candidate descriptions.

4. Compute description scores and select the highest scoring description.

In order to score descriptions in our domain (weather), we followed two intuitions. First that there is a minimum ratio of true positives a description can capture in order to be accepted as candidate. For example, if a description A overlaps with only 70% of the target points and description B with 90%, and we require at least 80% of true positives, description B is a candidate and A should be ignored. The second intuition states that, of all candidate descriptions, the description with the highest balance of true positives and true negatives should win. We used recall as the metrics for minimum threshold of true positives and F-measure as the metrics to balance out true positives and negatives. These are computed as (precision is also provided, since F-measure requires it):

$$precision = \frac{description \cap target}{description}$$

$$recall = \frac{description \cap target}{target}$$

$$Fmeasure = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Where *description* is the set of points associated with a description (e.g. (NORTH ⊓ COASTAL)) and *target* is the set of points associated with the target subset (e.g. those that represent rain).

Below is an example of the procedure with a hypothetical data set and target. Let us assume Figure 2a is the entire data set and represents the entire region, where the subset {C1, D1, E1, F1, E2, F2} is the target subset for which a description needs to

[3] The algorithm rejects intersections that are semantically redundant (e.g. ((NORTH, MORAY) ⊓ (MORAY)) ≡ (NORTH, MORAY)) or linguistically awkward (e.g. ((NORTH, MORAY) ⊓ (NORTH) → "the area of intersection between the North of Moray and the North of the whole region").

be generated. The preparatory step before the algorithm starts is to do geocharacterization with the absolute Frames of Reference. Let us assume our full geocharacterized model should contain 3 frames: NamedAreas, COASTAL and Directions. Directions is a relative frame and needs the descriptors of NamedAreas to exist, so initially we can only construct the frames CoastalProximity and NamedAreas (Figures 2b and 2c).

At any given point, a description is only considered as candidate if it scores higher than 0 recall, i.e. if it intersects at least once with the target set. This results in the following initial list of candidate descriptions (where R=recall and F-M=F-measure):

Absolute Descriptions	**R**	**F-M**
COASTAL	0.83	0.59
ABERDEENSHIRE	1.00	0.40

Now the algorithm creates mereological descriptions with the Directions frame (Figure 2d), as explained in Section 2. Once this interim geocharacterization step is done, mereological descriptions are added to the list of candidates:

Mereological Descriptions	**R**	**F-M**
NORTHEAST, ABERDEENSHIRE	0.83	0.67
NORTH, ABERDEENSHIRE	1.00	0.67
EAST, ABERDEENSHIRE	0.83	0.45
NORTHWEST, ABERDEENSHIRE	0.17	0.22
EAST, ABERDEENSHIRE	0.17	0.13

The next step is to generate intersections between all current candidate descriptions, as long as they are valid (see above), and add them to the list of candidates:

Intersected Descriptions	**R**	**F-M**
COASTAL ⊓ (NORTH, ABERDEENSHIRE)	0.83	0.83
COASTAL ⊓ (EAST, ABERDEENSHIRE)	0.83	0.77
COASTAL ⊓ (NORTHEAST, ABERDEENSHIRE)	0.67	0.73
COASTAL ⊓ ABERDEENSHIRE	0.83	0.71
COASTAL ⊓ (NORTHEAST, ABERDEENSHIRE)	0.17	0.29

Once the overgeneration algorithm is done, the scoring algorithm chooses the description with highest F-measure score, after filtering by recall. Assuming a recall threshold of 0.80, the description (COASTAL ⊓ (NORTH, ABERDEENSHIRE)) is the winner, as it has the highest F-Measure score of all remaining candidates. However if there is a need to raise the recall threshold to 1.00, i.e. no target point must be ignored, then the winning description is (NORTH, ABERDEENSHIRE). The choice for a particular recall threshold may vary from domain to domain. In the studies we have carried out, we achieve best performance at a threshold of 0.60 for one testbed, and 0.80 for another, as explained in Section 5.

4 Knowledge Acquisition

In this section we explain how we created a corpus of aligned data and text, which had a two-fold use: (a) inform us about the spatial operations employed by humans when producing geographic expression, and (b) serve as a testbed to evaluate the development of the algorithm.

From the work of de Oliveira et al. (2015) it became evident that named areas played an important role in geographic referring expressions, especially by allowing a mereological relation between certain unnamed descriptors and named descriptors. However that study provided only a high-level understanding of how often each Frame of Reference is used by humans when producing geographic referring expressions. In this study we conducted an experiment to produce an aligned data-and-text corpus, where each expression is associated with a particular subset of points (similar to the SUMTIME-METEO corpus (Sripada et al., 2002)). This enables the use of corpus entries as test cases, by running the algorithm with the subset of points of each entry, and comparing the output of the algorithm with the description in the entry.

Another interesting aspect of the corpus is its source. The texts were written by human experts (2 meteorologists), which guarantees that the geographic expressions in the corpus are similar to those in published weather forecasts. We could not guarantee this if the same texts were written by non-experts, for example using crowd-sourcing platforms. Nonetheless it is important to remember that our corpus – as strongly advised by the experiment participants – does not reflect the nature of real-life weather reports, with all the complexity that is involved in describing the weather. The corpus we present here is a collection of geographic expressions written by people with a life-time experience in producing geographic expressions; it is not a collection of real-life-like weather reports.

Using a web-based tool[4], the experts were exposed to 20 data sets. Each data set hypothetically represented a simplified weather forecast for the Scottish Grampian Region. When plotted onto the map, data points that represented some form of precipitation were highlighted in red, as shown in Figure 3. The experts were asked to write a pseudo weather forecast, describing where precipitation and/or dry weather was expected.

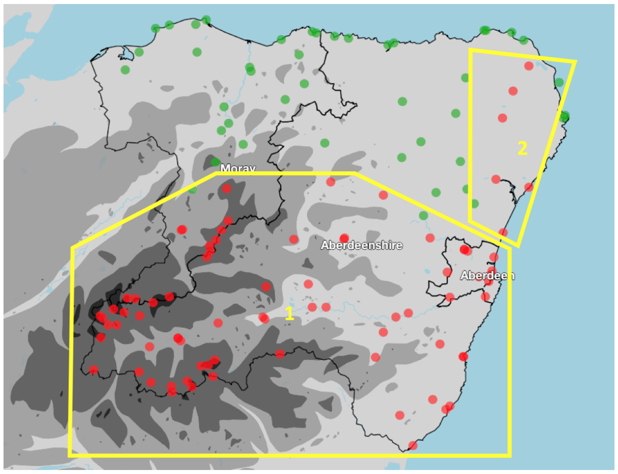


Figure 3: A map the meteorologists saw to write a weather forecast. Red points denote precipitation and green points dry weather. The numbered boxes were added for the alignment step, after texts had been written. Numbers on texts and boxes mark the alignment between points and expressions.

The above was only the first task of the experiment. The outcome of the the first task was a set of free-text paragraphs describing the location of wet and/or dry weather for the entire data set seen. The first observation we made from the raw responses is that some data clustering was taking place, because paragraphs contained many expressions (effectively noun phrases) to describe a single data set. This meant an alignment between parts of the texts and subsets (or clusters) of points had to be made. We prepared a document by hand where we provided the authors with screenshots of the maps they saw, along with the texts they wrote for each map. We numbered each expression on the texts and placed numbered boxes on the subset of points we judged to be referred to by each expression, as shown in Figure 3. The authors' task was to review our suggested alignment and fix it where applicable.

The last task to effectively build a corpus of data-and-text alignments was to annotate each referring expression with semantic labels. This task was carried out by one group of 3 human annotators per meteorologist – henceforth M1 and M2 – whereby 1 annotator participated in both annotations. The annotation task (for both M1 and M2) consisted of tagging expressions with labels of various categories. The following categories and labels were available:

Main direction Included the cardinal and intercardinal directions.

Direction modifier For words such as *far* and *central*, as well as the cardinal directions of complex direction expressions such as "NNW", where we assume the main direction to be NORTHEAST and the modifier to be NORTH-. This category is mainly for completeness, since we did not implement any of them.

Area The 3 Authority Areas of the Scottish Grampian region: ABERDEEN, ABERDEENSHIRE and MORAY.

Coastness Whether COASTAL or INLAND.

Altitude Whether HIGH or LOW.

Each category relates to a frame of reference in our system, and labels relate to descriptors. For each category, a null annotation was also available, in case the frame of reference was not mentioned. Annotators were instructed to annotate expressions entirely based on the linguistic material provided, not using their world knowledge. For example, if they were familiar with Aberdeen City and recognized it as a coastal city, but the expression was simply "Aberdeen", they should provide only { ABERDEEN } as annotation and not { ABERDEEN, COASTAL }.

Overall agreement between annotators was high – 92% for M1 and 98% for M2 – whereby the category Coastness had the highest disagreement (63%)

[4] http://homepages.abdn.ac.uk/rodrigodeoliveira/pages/georef/index_ka.php

for M1, as shown in Table 1. This was probably due to bad instructions as we suspect one annotator was using his world knowledge to judge whether a referred area was close to or far from the Grampian coast. All annotators live in Aberdeen City, but they saw only the expressions and no images. We improved instructions before annotating M2.

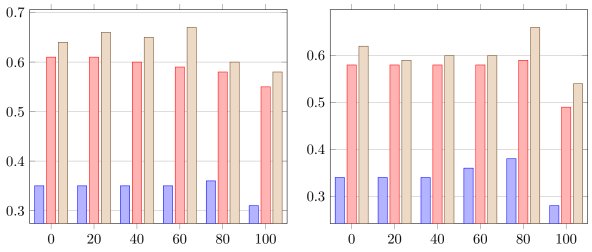
M1 sub-corpus	**AB**	**AC**	**BC**	**ABC**
Main direction	1.00	0.98	0.98	0.97
Direction modifier	1.00	0.96	0.96	0.97
Area	1.00	1.00	1.00	1.00
Coastness	0.92	0.52	0.46	0.63
Altitude	1.00	1.00	1.00	1.00
All categories	**0.98**	**0.89**	**0.88**	**0.92**
M2 sub-corpus	**AD**	**AE**	**DE**	**ADE**
Main direction	1.00	0.97	0.97	0.98
Direction modifier	0.96	0.87	0.91	0.92
Area	1.00	1.00	1.00	1.00
Coastness	1.00	1.00	1.00	1.00
Altitude	1.00	1.00	1.00	1.00
All categories	**0.99**	**0.97**	**0.98**	**0.98**

Table 1: The Kappa agreement scores when labelling expressions produced by both meteorologist (M1 and M2). Columns 2-4 show the pair-wise agreement, and the column 5 the averages of pairwise agreements per category. Figures at the bottom of each sub-corpus are the averages of each column.

After annotation, there were no cases where all three annotations were different, so there was a most frequent annotation for each data set. We kept those as the final set of labels for each entry in the corpus. After annotation, the M1 sub-corpus contained a total of 57 data-and-text aligned entries, while M2 contained 41. In the next section we explain how we used both M1 and M2 to evaluate the progress when developing the algorithm.

5 Evaluation and Discussion

Our algorithm development was carried out in two phases. First, we used a Gold Standard from M1 to develop the logic of the algorithm, and subsequently used a Gold Standard from M2 to test its performance. For each phase, we ran the algorithm with 3 distinct combinations of spatial operations: a) no operation, so only absolute descriptions such as (COASTAL) and non-specific directions such as (NORTH) were generated; b) mereology only, where mereological descriptions such as (NORTH, MORAY) were generated in addition to the ones above; c) both mereology and intersection, where the most complex descriptions such as (COASTAL ⊓ (NORTH ⊓ MORAY)) were also generated. The evaluation method was intrinsic, as described by Belz and Gatt (2008), whereby we computed the similarity between corpus descriptions and the output of the algorithm using the DICE coefficient of similarity. The Gold Standard testbeds excluded descriptions with direction modifiers such as *far* and *central*, because the current algorithm does not have an implementation for these concepts. The Gold Standard from M2 contained 44 entries, and that from M2, 36.



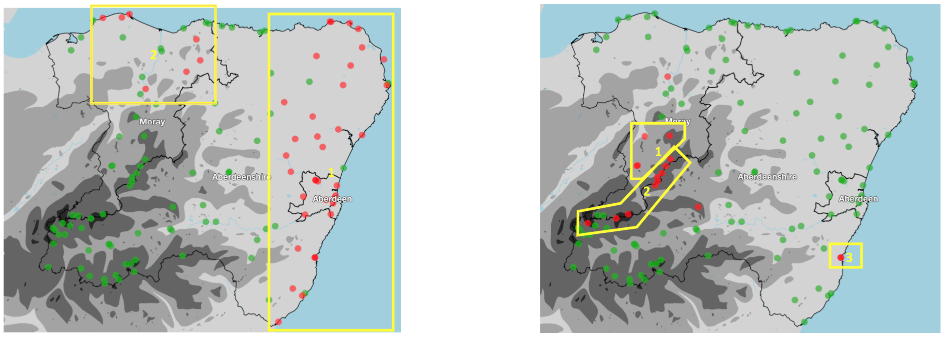
(a) Training scores (M1).

(b) Test scores (M2).

Figure 4: DICE similarity scores when running the algorithm against both sub-corpora (M1 and M2), using 3 different operation combinations – no operation (blue), mereology only (red), and both mereology and intersection (brown) – and 5 different recall thresholds. The X axis shows the different recall thresholds in percentage. The Y axis shows the average DICE scores across all data sets.

For each testbed we ran the algorithm 6 times, one for each recall threshold of an arbitrary set of thresholds (0.0, 0.2, 0.4, 0.6, 0.8 and 1.0). The results (shown in Figure 4) suggest that there is no specific recall threshold that gives better results, but 1.00 (i.e. no false positives accepted) is not the ideal threshold as it gave the worst results in all scenarios. However, the evaluation showed that there was a consistent gain in performance after the addition of each spatial operation. The highest average of DICE scores for M1 went from 0.36 with no operations to 0.67 with both operations, whereas for M2 scores went from 0.38 to 0.66.

We can attempt to explain why some of our output differs from the human descriptions. **Geocharacter-**



(a) (b)

Figure 5: Examples of almost perfect match between human-generated and machine-generated descriptions.

ization: If the mental models of the humans do not align with those our algorithm uses. An example of this is the description (EAST ⊓ COASTAL) which the human M2 gave to cluster 1 of the map in Figure 5a. The winning description according to the algorithm was only (EAST), because (EAST ⊓ COASTAL) covered less of the target points. This relates also to the topic of vagueness (Van Deemter, 2009), if one assumes descriptors not to have crisp but fuzzy boundaries (Schneider, 2000; Bittner and Smith, 2003). **Weighting:** If some descriptions should be rewarded if they include certain descriptors. This is much in line with the preference order of properties from the Incremental Algorithm (Dale and Reiter, 1995). The human-generated description for cluster 2 on Figure 5b was (HIGH ⊓ (SOUTH, MORAY)) which was the second best description generated by the machine. If the algorithm rewarded descriptions that include a named area, maybe the above description would have won.

These are only some of the possible reasons. We may not forget either that discourse and brevity may also play a role. Nonetheless the results we present in this paper show how, in any scenario, an algorithm for generating geographic expressions performs better if it employs intersection and mereology than without any operation.

6 Conclusions and Future Work

In this paper we have outlined an algorithm for generating geographic referring expressions. The algorithm employs 2 spatial operations – intersection and mereology – when processing point-based data. We described the compilation of a data-and-text aligned corpus, which we used as a testbed to guide development and to test the final system. We have shown that employing spatial operations makes the machine-generated output more similar to the human-generated descriptions. We increased the overall average of similarity between the computer output and human descriptions from a 0.38 (DICE), when no operations are used, to a score of 0.66, when computing mereology and intersection.

In line with Reiter and Belz (2009), we believe that our metrics-based evaluation was valuable but only a ‘development-stage’ guidance. A task-based evaluation shall be more revealing of the algorithm’s performance. Thus, our next study will evaluate how well users accomplish a task given the descriptions generated by our algorithm. Nonetheless we are convinced that spatial operations are employed by humans when producing descriptions, which makes the algorithm described here to be more human-like than previous approaches. Above all, our results show that relative properties are paramount when generating referring expressions in geographic domains, where mereological relations are key.

7 Acknowledgements

We would like to thank Arria NLG for sponsoring this PhD project. Our special thanks are due to the meteorologists Mr Ian Davy and Mr James Brownhill for providing examples of spatial referring expressions which guided the development of our algorithm. Many thanks to the members of the Computational Linguistics Aberdeen group for their helpful comments, especially Kees van Deemter, Steph Inglis and Daniel Couto-Vale.

References

John Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14):1027–1071.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 197–200. Association for Computational Linguistics.

Thomas Bittner and Barry Smith. 2003. Vague reference and approximating judgments. *Spatial Cognition & Computation*, 3(2-3):137–156.

Anthony G Cohn and Jochen Renz. 2008. Qualitative spatial representation and reasoning. *Handbook of knowledge representation*, 3:551–596.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Rodrigo de Oliveira, Somayajulu Sripada, and Ehud Reiter. 2015. Designing an algorithm for generating named spatial references. *ENLG 2015*, page 127.

Andrew U Frank. 1992. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing*, 3(4):343–371.

Albert Gatt, Roger PG van Gompel, Kees van Deemter, and Emiel Kramer. 2013. Are we bayesian referring expression generators. In *Proceedings of CogSci*, volume 35.

Helmut Horacek. 2004. On referring to sets of objects naturally. In *Natural Language Generation*, pages 70–79. Springer.

John D Kelleher and Geert-Jan M Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1041–1048. Association for Computational Linguistics.

Imtiaz Hussain Khan, Kees Van Deemter, and Graeme Ritchie. 2008. Generation of referring expressions: Managing structural ambiguities. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 433–440. Association for Computational Linguistics.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Stephen C Levinson. 2003. Space in Language and Cognition: Explorations in Cognitive Diversity. chapter 2, pages 24–61.

Inderjeet Mani, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. 2010. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280.

David M Mark, Christian Freksa, Stephen C Hirtle, Robert Lloyd, and Barbara Tversky. 1999. Cognitive models of geographical space. *International journal of geographical information science*, 13(8):747–774.

Alejandro Ramos-Soto, Nava Tintarev, Reiter Ehud de Oliveira, Rodrigo, and Kees van Deemter. 2016. Natural language generation and fuzzy sets: An exploratory study on geographical referring expression generation. In *Proceedings of Fuzz-IEEE 2016*. IEEE.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*, volume 33. MIT Press.

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

Markus Schneider. 2000. Finite resolution crisp and fuzzy spatial objects. In *Int. Symp. on Spatial Data Handling*, page 5a. Citeseer.

Somayajulu Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2002. Sumtime-meteo: Parallel corpus of naturally occurring forecast texts and weather data. *Computing Science Department, University of Aberdeen, Aberdeen, Scotland, Tech. Rep. AUCS/TR0201*.

Ross Turner, Somayajulu Sripada, and Ehud Reiter. 2010. Generating approximate geographic descriptions. In *Empirical methods in natural language generation*, pages 121–140. Springer.

Kees Van Deemter. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

Kees Van Deemter. 2009. Utility and language generation: The case of vagueness. *Journal of Philosophical Logic*, 38(6):607–632.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Michael F Worboys and Matt Duckham. 2004. *GIS: a computing perspective*. CRC press.