Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions

Niloofar Aflaki
Massey University, Auckland, New Zealand
n.aflaki@massey.ac.nz

Shaun Russell
Massey University, Auckland, New Zealand
shaun@ensemblemusic.co.nz

Kristin Stock
Massey University, Auckland, New Zealand
k.stock@massey.ac.nz

Abstract

In order to extract and map location information from natural language descriptions, a first step is to identify different language elements within the descriptions. In this paper, we describe a method and discuss the challenges faced in creating an annotated set of geospatial natural language descriptions using manual tagging, with the purpose of supporting validation and machine learning approaches to annotation and text interpretation.

2012 ACM Subject Classification Applied computing → Annotation

Keywords and phrases Annotation challenges, spatial relations, spatial language

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.20

Category Short Paper

Funding This work is partly funded through a Ordnance Survey PhD scholarship.

1 Introduction and Background

To progress research on the interpretation of geospatial natural language, methods for automated tagging of spatial language are required [5, 9]. In this paper, we discuss the challenges that we encountered when trying to create manually tagged annotated data set that addresses the shortcomings of previous data sets, using two experiments. A number of researchers have addressed the problem of annotating geospatial natural language. For example, Stock and Yousaf [10] annotated a wide range of language elements, including adverb and parts of objects as well as relatum, locatum and spatial relation, mainly by extending POS tags in a rule-based approach. Kordjamshidi et al [5] restrict their attention to trajector, landmark and spatial prepositions, although they acknowledge that other parts of speech can be used to express spatial relations. GUM Space specifies a broad range of tags including locatum, relatum, spatial modality [3]. SpatialML uses mark-up language to tag elements [7] including places, coordinate, orientations, form of reference, direction, distance and frame. Work by Zwarts [12] and Kracht [6] address spatial prepositions, with a focus on directional prepositions and location. Much of the previous work is either limited to very simple elements [5]; adopts a complex tag structure [3] or assumes a particular syntactic (grammatical) structure [5, 10]. We propose an annotation scheme that addresses these limitations in that it focuses on semantics rather than syntax.

2 Methodology

We conduct our exploration of the challenges of creating an annotated data set using two experiments. The first one compares the tagging conducted by pairs of human annotators and discusses discrepancies and issues involved in manual tagging. The second one discusses variations between individual human respondents in matching natural language descriptions to spatial relations, highlighting the lack of consensus.

2.1 Experiment 1: Creating an Annotated Data Set

The selection of an annotation scheme was based on three criteria: 1. What must be individually identified in order to support effective geocoding of the text? This is difficult to evaluate conclusively, as it depends upon the geocoding approach, and some aspects of spatial language are still not well understood. This criterion influences not only which items we tag, but also which items we identify as separate elements. For example, it is not useful to separate *next to* into two separate tags, because the meaning depends on the combination of the words, and the meaning of *to* in particular is dependent on the presence of *next*. In contrast, adverbs like *right*, or *directly*, have their own meanings which are similar regardless of the preposition they appears with, although the meaning may be influenced by the latter. 2. Can some of the tags or their subcategories be reliably determined automatically? If a particular semantic tag can be reliably identified through an automated approach, then there is little point in annotating in manually. The reliability of an automated approach is a question of degree, but we use the yardstick that if the set of words of interest can be defined by a clear set of specific words, none of which are homonyms, then they might reliably be identified automatically. In practice this is rare, because for example, even though the set of prepositions is a closed word class, since we are interested in semantic tags rather than syntactic, and prepositions normally encode spatial relations, there are examples of spatial relations that are not prepositions *(e.g.in line with)*. 3. What is practical to expect people to reliably annotate? This involves both volume and simplicity. A set of tags that is too complex will be difficult for manual annotators to deal with. The set of tags must be manageable in quantity, and simple enough to understand without specialist knowledge.

In Experiment 1, we develop a generic spatial annotation framework based on the semantic roles of tokens in a sentence. To this end, 1000 sentences were randomly selected from the combined set of three data sources: The Nottingham Corpus of Spatial Language[9], The Landcare Research National Soils Database [1] and The Where Am I survey, in which natural language descriptions were elicited from human respondents, as described in [8]. Table 1 identifies, describes and explains the annotation scheme that was used. Four annotators were given an expanded version of Table 1 with a simple explanation of terms and examples. The purpose of the work was explained to them in simple terms, and they were given access to the tagging app. Each annotator was then asked to annotate 10 expressions using the tagging app, after which the authors examined the expressions and gave feedback on any issues, before the annotator began annotating in earnest. Each expression was tagged twice by two different annotators.

[1] `https://soils.landcareresearch.co.nz/index.php/soil-data/national-soils-data-repository-and-the-national-soils-database/`

Table 1 Tag labels and descriptions.

Title	Explanation
Trajector	The object whose location is being described. The important role of the trajector in spatial language has been discussed by a number of researchers and is also known as locatum [3] or figure [11].
Landmark	The object that is used as a reference point in the description. The landmark also plays an important and well documented role in spatial language, and is similar to the relatum and ground identified by other researchers[11].
Spatial Relation	The word or words that indicate how two objects are positioned relative to other. The importance of spatial relations has also been well recognised, and they have been widely researched [1, 4, 12]. In syntactic terms, spatial relations are most often represented using prepositions, but not always.
Location and movement verb (lmv)	A verb that describes the manner in which one object is positioned relative to the other. The location and movement verb is a subset of the verb syntactic category[11]. *The road **crosses** behind the church.*
Spatial qualifier	A word of set of words that adds more information to the spatial relation and or the location and movement verb. Spatial qualifiers have not been widely recognized as an important carrier of spatial information as yet, and may be represented with a range of different parts of speech, including adverbs, adjectives and nouns. *The road goes **right** beside the church*
Spatial specifier	A word of set of words that describes particular subparts of a feature.E.g. *The **north of** the country.* Spatial specifiers have also not been widely studied in specific terms, with work instead focusing on general issues of mereology [2].

2.2 Experiment 2: Matching of Expressions to Spatial Relations

In the second experiment, we used data collected in earlier work [10]. In this work, respondents were shown expressions one at a time, and asked to match each expression to one of a series of diagrams that illustrated spatial relations. After viewing the expression and the set of available spatial relation diagrams, each annotator was asked to select values on a Likert scale that included only the positive side of the scale, to indicate his or her opinion about how closely each of the selected spatial relation diagrams matched the expression: *Strongly Agree, Agree, Agree Somewhat.* Only the positive half of the scale was used because users were invited to only select diagrams that they thought reflected the expressions (i.e. if they did not agree, the respective diagram would not be selected). Weights were allocated to each response for a given spatial relation diagram-expression pair, using 1, 0.75 and 0.5 for Strongly agree, Agree and Agree Somewhat respectively. The score for each expression and its geometric configuration was calculated using this formula:

$$GCOScore_{expression,diagram} = \sum_{k=0}^{n}(response_k weight_k)/n \quad (1)$$

In which *response k* represents the number of responses with *weight k*, and n defines the total number of responses for expression k. Full details of the methodology can be found in[10].

3 Results

In order to evaluate the reliability of the manual annotation process in Experiment 1, we calculate inter-annotator agreement among the four annotators. Since expressions were randomly allocated to annotator, any combination of pairs of specific annotators may annotate a given expression. Inter-annotator agreement was calculated by comparing the words in

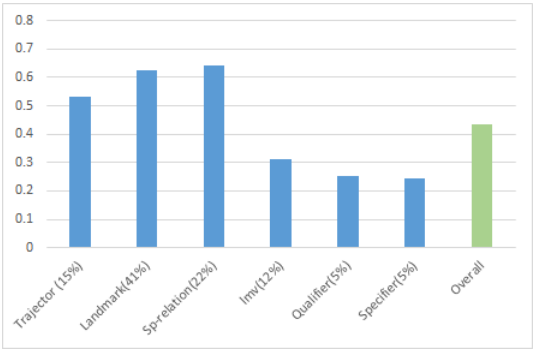

Figure 1 Study 1. Mean inter-annotator agreement by tag type.

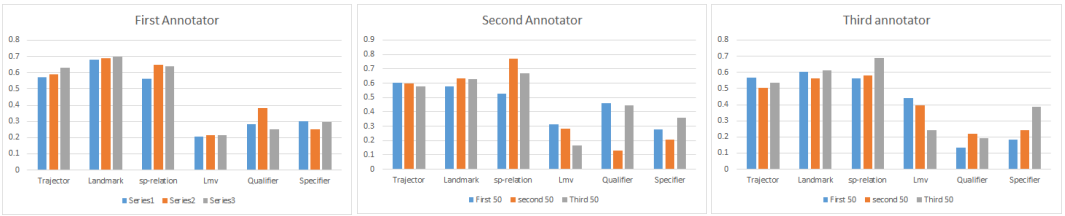

Figure 2 a-c Annotator performance through the time.

a given expression that were given a particular tag by each annotator. Since many of the expressions were complex and contained more than one of some tags, we calculate agreement by proportion of overlap between the words annotated with a particular tag by each user, rather than by a simple true/false agreement. Formula 2 expresses this measurement of agreement between annotators for a single expression: For a given tag, ME_k denotes the number of mutual elements (words or multi-word tagged values) that both annotators agree on, and max_k denotes the maximum number of elements that are tagged by either annotator. The total agreement score for the expression is then average of agreement across the populated tags. For example, if user 1 specifies Australia, New Zealand and Canada as landmarks and user 2 specifies Canada and USA as landmarks ME_k for the landmarks would be 1, because just Canada is mutual and the max_k would be three as the maximum number of landmarks by either annotator. The agreement score is calculated for all the tags in an expression, and the average is calculated to determine the agreement across the entire expression.

$$AgreementScore = Average(\sum(ME_k/max_k)) \tag{2}$$

Figure 1 shows the mean inter-annotator agreement for individual tags, as well as overall and also the percentage of tags of each type that were annotated in the 1000 expressions. We used this formula, to have an accurate calculation of each separate tag. We also explore the role of annotator experience in the manual tagging process, and evaluate whether annotator performance improves over time. For each annotator, we calculated inter-annotator agreement for the first, second and third 50 expressions tagged by three annotators through the time to see whether their performance changed by time or not. Only 3 annotators are shown because the remaining did not annotate sufficient expressions. Figures 2a to c show the results. We then calculated the inter-annotator agreement of different subsets of annotators, to determine whether some annotators were more successful than others in tagging, either overall of for specific tags. The results (Figure 3) show some inconsistency. It is, however, clear that Annotator 2's contribution is important, with her exclusion resulting in overall deterioration.

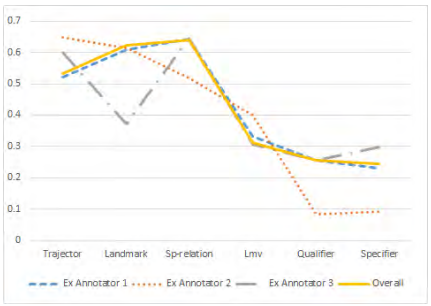

Figure 3 Inter-annotator agreement excluding each annotator in turn.

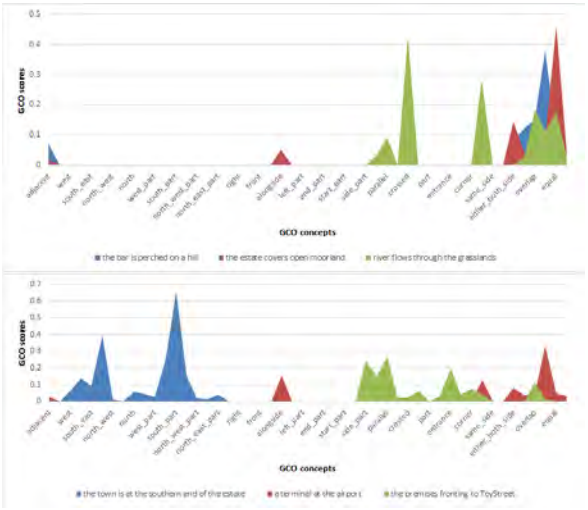

Figure 4 a,b Study 2. GCO score for second three expression.

Turning to Experiment 2, the results highlight the lack of agreement among individual respondents regarding the spatial relation diagram that best reflects a given expression. The respondents in Experiment 2 were also non experts in geographic information science. Figures 4 a and b each show the spread of responses for three example expressions. In contrast to Experiment 1, Experiment 2 used short, simple spatial expressions, and the graph shows the frequency (after weights have been applied as described in Section 2) of selection of each spatial relation for a given expression. Two expressions in 4b show a number of small peaks, with no clearly dominant relation selected by the respondents. Across the entire data set, a similar pattern was observed, with lack of consensus among respondents in selecting spatial relations to match many expressions.

4 Discussion and Conclusion

The results clearly show that it is not straightforward to create a manually annotated data set of natural language descriptions with a broad set of language elements that is based on semantics rather than syntax. Obviously, for an annotated data set for use in machine learning and validation, we would like the agreement to be very strong. Considerations of the level of experience of the annotators and the examination of the influence of specific annotators on particular tags did not result in noticeable improvement. The challenges that were encountered can be summarised as follows: Firstly, it is not unusual for the same place name, geographic feature or moving object to be both a trajector and a landmark, and secondly, the landmark/trajector status of a word may be ambiguous. The following example illustrates both of these cases. In the expression *the church stands beside the post office near the bridge*, the structure of the expression could be:

"trajector+(lmv)+spatial-relation+landmark+spatial relation+landmark"

"trajector+(lmv)+spatial-relation+(trajector and landmark)+spatial relation+landmark"

In the first case, church is a trajector for both the church landmark and the bridge landmark, and in the second case post office is the trajector for the bridge landmark, as well as the landmark for the church trajector. The annotation scheme used in this paper allowed each word to be tagged only as a trajector or a landmark, but not both. The creation of a

tag that indicates a dual role may be a possible methods for addressing this. Resolution of ambiguity is a more difficult problem to solve, and even the most expert and experienced annotators may disagree. A final observation from the results is that, spatial qualifiers and spatial specifiers had only fair inter-annotator agreement (lower than other tags), and while this may be in part due to confusion about when to use each, when questioned, Annotator 2 was able to accurately explain when the spatial specifier tag was used and claimed to find it easy to understand. Confusion in the tagging process was sometimes caused by considerations of grammar, rather than meaning.

In this paper, we have described a semantic annotation scheme that is designed to be both useful and practical, and the methodology used to create an annotated data set. We analysed and presented some of the challenges encountered in the process, and the fundamental difficulties resulting from ambiguity and individual discrepancies in the use of spatial language that make it difficult to define a single, reliable annotated data set at a semantic level. In future work, we intend to do more analysis and test different annotation strategies like single tag per annotator, to see if there is any improvement in the results achieved.

References

1 Kenny R Coventry and Simon C Garrod. *Saying, seeing and acting: The psychological semantics of spatial prepositions.* Psychology Press, 2004.

2 Torsten Hahmann and Michael Gruninger. A naive theory of dimension for qualitative spatial relations. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

3 R Ross J Bateman J Hois, T Tenbrink, R Ross, and J Bateman. Gum-space. Technical report, Technical report, Universität Bremen SFB/TR8 Spatial Cognition, 2009.

4 John D Kelleher and Fintan J Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, 2009.

5 Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):4, 2011.

6 Marcus Kracht. The fine structure of spatial expressions. *Syntax and semantics of spatial P*, pages 35–62, 2008.

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

8 Kristin Stock, Didier Leibovici, Luciene Delazari, and Roberto Santos. Discovering order in chaos: using a heuristic ontology to derive spatio-temporal sequences for cadastral data. *Spatial Cognition & Computation*, 15(2):115–141, 2015.

9 Kristin Stock, Robert C Pasley, Zoe Gardner, Paul Brindley, Jeremy Morley, and Claudia Cialone. Creating a corpus of geospatial natural language. In *International Conference on Spatial Information Theory*, pages 279–298. Springer, 2013.

10 Kristin Stock and Javid Yousaf. Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *International Journal of Geographical Information Science*, pages 1–30, 2018.

11 Leonard Talmy. *Toward a cognitive semantics*, volume 2. MIT press, 2000.

12 Joost Zwarts. Prepositional aspect and the algebra of paths. *Linguistics and Philosophy*, 28(6):739–779, 2005.