Component-wise Annotation and Analysis of Informal Place Descriptions

Igor Tytyk, Tim Baldwin

Department of Computing and Information Systems
The University of Melbourne
Melbourne VIC 3010, Australia
ihor.tytyk@gmail.com, tb@ldwin.net

Abstract. We analyse the strategies used in formulating situated informal location descriptions, by identifying geospatial expressions contained therein and annotating each for properties such as geospatial granularity and identifiability. Analysis of the annotations leads to insights such as the predominance of suburb-level expressions, and prevalence of vernacular expressions.

Key words: Informal place description, geospatial expression, named entity, vernacular geography, computational linguistics

1 Introduction

When informally describing one's whereabouts or giving directions, people make heavy use of place descriptions. In the descriptions they relate their location to the surrounding objects, or landmarks [1]. In order to make the instructions interpretable by the recipient, the description provider should use familiar landmarks and relate the location to them appropriately. Thus, for a human recipient this task is trivial. However, computational systems cannot easily interpret place descriptions expressed in natural language, or generate natural-sounding route or place descriptions.

Additionally, humans frequently make use of vernacular place descriptions, or refer to landmarks using non-standard renderings of their 'official' names, as a result, making it hard for computers to understand the description, and also humans unfamiliar with the locality being described. Wu and Winter state that placenames and spatial relations are main components of place descriptions, and in order to interpret the descriptions their components must be interpretable [2].

In this study we focus on analyzing placenames in the context of informal place descriptions, that is placenames that are elicited naturally and in situ, without any constraints or guidance. We manually identify geospatial expressions in a dataset of placename descriptions, and further annotate the granularity level, identifiability and normalised name of each such expression.

2 Dataset

Winter et al. collected situated place descriptions from players of the *Tell us where* location-based mobile game [3].[1] The game consisted of submitting textual descriptions of the location of smart phone users, along with their GPS location. The reasons we chose to use this data are many fold. First, the data was collected across a broad sample of users, ensuring the heterogeneity of the data and reducing sample bias. Second, the participants were asked to submit textual descriptions of their location from anywhere in the state of Victoria, Australia. This led to a diversity of locations, but within a restricted area of familiarity to our annotators and with the expectation of consistency in the strategies used by the participants to describe their location. Third, the users were given no guidelines for writing the descriptions, meaning that the data is rich in vernacular placename descriptions and the strategies used by users to describe their location are varied. Lastly, since the participants were using their mobile phones and basing the placename descriptions on their actual location. As a result, the descriptions are situated, spontaneous, and as natural as we could hope for.

A total of 2221 place descriptions were collected through the *Tell us where* game. However, the data contained duplicates. Since we are interested in qualitative rather than quantitative data, it was decided to eliminate all duplicates from the corpus. As a result, the final number of descriptions was 1858.

2.1 Annotation

We manually annotated the placename descriptions for *geospatial expressions*, in the form of: (1) geospatial named entities (*Federation Square, Swanston Street*); and (2) geospatial noun phrases (*school, a leafy park*). Named entities are proper names, and are generally subclassified according to the semantic class of the referent, e.g. into persons, locations and organisations. However, for the purposes of this research, we restrict our attention to geospatial named entities.

One of the broader goals of this work is the compositional semantic interpretation of place descriptions. It was thus decided that we should aim for maximum segmentation granularity in our annotation, while avoiding nested annotations. For example, if the place description were an address such as *Melbourne University Bookshop, in Parkville near the library*, we would segment it into the geospatial named entities *Melbourne University Bookshop* and *Parkville*, and the geospatial noun phrase *the library*. Note that we would not also identify *Melbourne University* as a geospatial named entity, as it is nested within *Melbourne University Bookshop*.

We expected many of the geospatial expressions in the dataset to be noun chunks. For example, *Queen Victoria Market* is a single noun chunk geospatial named entity, while *a tall building* is a single noun chunk geospatial noun phrase

[1] `http://telluswhere.net/`

Granularity level	Description
(1) Furniture	Location within a room, referring to furniture (*by my computer, in bed*)
(2) Room	Location within a building, or parts belonging to it (*in my room, third floor*), or medium-sized vehicles (*car, train*)
(3) Building	Location of a building, street no. or building name (*geomatics dpt, street corner/intersection,*)
(4) Street	Institution, public space or street level, larger than building and/or vaguer boundaries than building. For example, transport infrastructure (*railway, tramline, Ave, Circuit*), a public space (*school, cemetery, mall*), or a natural landmark (*lake, park*)
(5) District	Suburb, rural district or locality, or post code area (*carlton, North Melbourne, CBD*)
(6) City	Town or city level, and metropolitan areas (*Canberra, near Geelong*)
(7) Country	Everything beyond city level, including highways, freeways (*Princes Hwy*), islands (*French Island*), rivers (*Murray river*) and states (*WA*)

Table 1. Granularity level classification (Richter et al., 2012); all examples are taken from the actual dataset, and are presented using the original orthography

referring to a construction, which can be used as a reference point when describing a location. In the interests of expediting annotation, we first chunk-parsed the place descriptions, using the Stanford CoreNLP tools.

The annotation scheme we used is comprised of several layers. The first annotation layer contains information about whether a given segment is a geospatial named entity (*NE_NP*) or a geospatial noun phrase (*NP_NP*). The remaining layers apply to each geospatial expression.

The second layer of annotation is the *granularity level*, and captures the "zoom level" of each geospatial expression. The granularity level is judged on the scale from 1 to 7, based on the classification of Richter[4] as detailed in Table 1. In some instances, we diverge from Richter's classification. For example, when a named entity is too big or too small for the bounding box of its default zoom level, we override the default to capture the zoom level which best matches the size of the bounding box. *Mountain Highway*, e.g., goes through only a few suburbs of Melbourne, so we override the *Country* granularity level for highways and assign it to the zoom level of *City* to better reflect its size. Similarly, when determining the granularity level of towns, it was decided to shift the small towns that do not have suburbs (e.g. *Warragul* and *Pakenham*) from *City* to *District*.

The third layer of annotation is *identifiability*. This captures whether a geospatial expression is unique within Victoria or there are multiple instances of it. There are three possible values for identifiability: *non-identifiable*, *identifiable ambiguous*, and *identifiable non-ambiguous*. All geospatial noun phrases (e.g. *school, park, monument*) are non-identifiable, since the set of these objects within Victoria is very large and it is not possible to geocode them without disambiguating information. Some geospatial named entities are considered to be non-identifiable due to their ubiquity and unavailability within standard gazetteers of an exhaustive listing of every instance within Victoria (e.g. *Mc-*

Donalds, 7-eleven). On the other hand, a geospatial named entity can refer to a small set of several places which are enumerated in a gazetteer, in which cause they are considered to be identifiable ambiguous. For example, there are four instances of *Canning Street* in Victoria, so every *Canning Street* in the corpus is annotated as identifiable and ambiguous. On the other hand, *Flemington Road* is identifiable non-ambiguous as there is only one instance in Victoria.

As with granularity, the determination of identifiability is inevitably subjective. To reduce the effects of subjectivity as much as possible, we base the judgement on two online gazetteers: OpenStreetMap[2] and Google Maps.[3]. Google Maps contains an extensive listing of named entities, but has poor coverage over non-standard or vernacular equivalents of less well-known named entities. Thus, while *melb uni* (standard = *The University of Melbourne*) and *fed square* (standard = *Federation Square*) can be found in Google Maps, it does not contain local vernacular such as *broady* (standard = *Broadmeadows*) or non-standard abbreviations such as *pi* for *Phillip Island* or *fg* for *Ferntree Gully*. Here, we elicited support from locals and the Google search engine to interpret the geospatial expression.

Names of cafes, restaurants, and other small businesses were the most difficult NEs to judge identifiability for. Even though OpenStreetMap lists a vast number of buildings, eating places, shops, many of them were missing.

The fourth and final level of annotation is the *placename normalisation*. Since the place descriptions were submitted by mobile phone, the dataset contains a lot of abbreviations, misspellings and vernacular names. The canonical name/spelling was provided in all such instances. For example, *melb uni* would be normalised to *The University of Melbourne*. We observed an inevitable dependency between identifiability and placename normalisation for geospatial named entities: if a geospatial named entity cannot be identified, it is not possible to determine its normalised rendering.

Some of the submitted place descriptions do not contain any information about the location (e.g. *this will be an everlasting love*) or are located outside of Victoria (e.g. *in Wagga Wagga*). All such descriptions were marked as irrelevant at the message level, using the *IRREL* label.

For the annotation we used brat,[4] a highly-configurable, easy-to-use web-based text annotation tool.

3 Analysis and Discussion

Having annotated the dataset, we extracted a feature vector for every annotated geospatial expression (excluding the irrelevant descriptions). Each feature vector contained a set of values: *id*, geospatial expression *type*, *granularity level*, *identifiability*, *original* spelling, and *canonic* (normalized) spelling. Then, all the vectors

[2] `http://www.openstreetmap.org`
[3] `http://maps.google.com.au/`
[4] `http://brat.nlplab.org/`

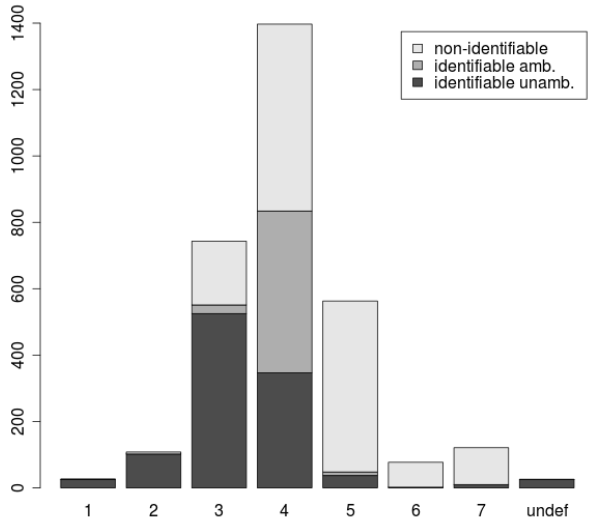

Fig. 1. Granularity level vs. identifiability in the dataset

were collated into a table and fed into the R statistical package[5] for analysis. In total, 3061 geospatial expressions were extracted, 2139 (70%) of which were geospatial named entities. That is, without any constraint on the description, about two thirds of geospatial expressions contained in place descriptions can potentially be found in gazetteers.

Figure 1 presents a distribution of geospatial expressions across zoom levels, broken down by identifiability. The mean granularity value is 4.01, with a standard deviation of 1.05. The most common granularity level is 4 (*Street*), with about 45% of all geospatial expressions. This means that when writing place descriptions, users tend to make heavy use of streets, parks, squares, universities and hospitals. Of the remainder, almost a quarter (24%) of the referents are of the *Building* granularity level (level 3), and about 18% are of the (*Suburb*) granularity level (level 5).

The correlation between the granularity level and the fraction of non-identifiable placenames is not very surprising: the bigger the spatial feature, the more likely it will be identifiable. On the other hand, the appreciable drop in non-identifiability at the *Suburb* level is proof of the salience and unambiguity of the placenames within this level. After dividing all the geospatial expressions by identifiability and filtering out from the non-identifiable ones the names of chain stores and eating places (e.g., *McDonald's, Subway, Coles*), it is possible to calculate how many of the named entities are not in the gazetteers (Open-

[5] `http://www.r-project.org/`

StreetMap and Google Maps). Out of 2139 named entities, 51 (2.4%) are not contained in the gazetteers. As a rule, among these placenames are names of restaurants, apartment blocks, and other small scale companies (e.g. *Pilkington Glass, Ching Chong Food, Yarra Crest Appartments*).

Another important category of geospatial expression is vernacular descriptions. We found a considerable number of entrenched vernacular equivalents of salient Victorian placenames, and common strategies for forming vernacular place names. Some of them are formed by simply dropping one of the constituent words (*Narre Warren* → *narre*), some by "clipping" the word (*Yackandandah* → *yack*, *Dandenong* → *dande*), and some are acronyms (*Phillip Island* → *pi*, *Ferntree Gully* → *fg*). However, the most productive pattern was "embellished clipping", shortening the expression to the first syllable and adding a diminutive suffix *-y, -ie*, (e.g. *Richmond* → *richy*, *Beaconsfield* → *beacy*, *South Gippsland Highway* → *south gippy*). The pattern is particularly peculiar to the Australian English. From the collected informal NEs, one can infer that only salient and unambiguous placenames undergo the process of vernacularization. Since suburb names in Victoria are unique and widely used for describing locations, they are most commonly substituted by their informal equivalents.

4 Conclusions and Future Work

In this paper, we have performed detailed component-wise analysis of informal place descriptions. From this study, we can conclude the following: (a) most geospatial expressions are streetnames, parks, buildings and suburbs; (b) the presence of a suburb-level placename in the description increases its identifiability; (c) vernacular place descriptions are commonly used, based on a small number of strategies; and (d) geospatial named entities which are mostly likely to not be contained in gazetteers are names of pubs, cafes, and small businesses.

This paper has considered placenames independently of the message-level interpretation. A logical next step is a compositional analysis of the place description based on the annotations we have done, and investigation of how spatial relational semantics (e.g, prepositions like *near, at, in*) impacts on message interpretability and the properties of its constituent placenames.

References

1. A. Klippel. *Wayfinding Choremes Conceptualizing Wayfinding and Route Direction Elements*. PhD thesis, Universitaet Bremen, 2003.
2. Y. Wu and S. Winter. *Interpreting Destination Descriptions in a Cognitive Way*. Schloss Dagstuhl, Dagstuhl, 2011.
3. S. Winter, K.-F. Richter, T. Baldwin, L. Cavedon, L. Stirling, M. Duckham, A. Kealy, and A. Rajabifard. Location-based mobile games for spatial knowledge acquisition. In Janowicz et al., editor, *Cognitive Engineering for Mobile GIS*, Belfast, Maine, USA, 2011.
4. D. Richter, M. Vasardani, L. Stirling, K.-F. Richter, and S. Winter. Zooming in zooming out hierarchies in place descriptions. Unpublished manuscript, 2012.