Landmark Classification for Route Directions

Aidan Furlan†, Timothy Baldwin† and Alex Klippel‡

† CSSE
University of Melbourne
VIC 3010, Australia
{afurlan,tim}@csse.unimelb.edu.au

‡ Department of Geography
Penn State University
University Park, PA 16802, USA
klippel@psu.edu

Abstract

In order for automated navigation systems to operate effectively, the route instructions they produce must be clear, concise and easily understood by users. In order to incorporate a landmark within a coherent sentence, it is necessary to first understand how that landmark is conceptualised by travellers — whether it is perceived as point-like, line-like or area-like. This paper investigates the viability of automatically classifying the conceptualisation of landmarks relative to a given city context. We use web data to learn the default conceptualisation of those landmarks, crucially analysing preposition and verb collocations in the classification.

1 Introduction

At present, many navigation systems produce badly-worded and difficult to follow route instructions, which do not closely correspond with the way people give one another directions (Dale et al., 2005). Typically, automated navigation systems give turning instructions with street names as reference points, eg *turn right at Smith St*. By contrast, human-generated route instructions tend to use landmarks in preference to street names as navigational reference points (Michon and Denis, 2001).

According to Allen (1997), landmarks are typically used in route directions in one of two ways— as descriptives, providing a static picture of a spatial scene so that the traveller can verify his or her location along a route, eg *the City Library is on your left*, or to specify or clarify a point on a route at which the traveller must make a choice between multiple pathways, termed choice points or decision points. Route instructions which identify decision points with respect to landmarks have been found to be significantly easier to follow than standard street-based or turn-based route instructions (Michon and Denis, 2001).

This paper goes beyond classical approaches to landmarks that focus on salient point-like objects. Instead, we aim to find appropriate ways of classifying landmarks automatically, based on the way those landmarks are used in spatial sentences on the web: as point-like, linear-like, and area-like objects that structure movement pattern in urban spaces. In particular, we analyse how different prepositions and verbs with pre-classified semantics co-occur with mentions of the landmarks. A preposition such as *through* can be used with reference to a landmark we are conceptualising as an area, but not one we are conceptualising as a point. Landau and Jackendoff (1993) presented an analysis of the spatial properties of commonly used English spatial prepositions, such as *at*, *in* and *to*. This classification used as the basis of a list of prepositions for the present study, grouped according to whether the preposition indicates a point-like, line-like or area-like landmark. In addition, a short list of verbs was compiled based on the verb classes of Levin (1993) and similarly divided into the three conceptual classes.

Each of the verbs and prepositions was combined in turn with a list of landmarks in Melbourne, Australia, to produce a series of spatial phrases such as *at Flinders St Station*. These phrases were then

Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, pages 9–16,
Prague, Czech Republic, June 2007.

sent to the Google search engine, which determined the approximate number of documents on the web containing that exact phrase. The document counts were then summed over the conceptual categories the prepositions and verbs appeared in — point, line and area. The result of this was a probabilistic categorisation of each landmark, according to its usage in spatial contexts on the web.

Evaluation of the baseline was performed based on annotators' independent judgements of the conceptual class of each of the landmarks, gathered from a web-based annotation interface. It was found that the baseline classification agreed with the gold standard classification 63.8% of the time. A slight improvement on the baseline was achieved via a supervised neural network classifier, which took the web counts as inputs. This classifier agreed with the gold standard 68.5% of the time. As a result of this analysis, a set of systematically ambiguous landmarks was identified, with implications for future landmark classification models.

In the remainder of this paper, we describe background research (Section 2) and then outline our research methodology (Section 3). We then present the results of a series of landmark classification experiments (Section 4), and finally discuss the broader implications of the experiments (Section 5).

2 Background

2.1 Spatial Cognition

Route directions should be designed in such a way as to be quickly and easily comprehended by the traveller (Lovelace et al., 1999). Optimally, route directions should exhibit cognitive adequacy — characterising an external representation of a route (as with a map or route directions) in a way supportive of human spatial cognitive processes and knowledge representation (Klippel, 2003). For this reason, the improvement of route directions requires an investigation into human spatial cognition.

Route instructions which reference landmarks are able to achieve a number of worthwhile goals: they have the effect of increasing the spatial awareness of the recipient by informing them about their surroundings; landmark-referencing route instructions can decrease the cognitive load on the recipient; and it is more natural-sounding to receive route instructions in terms of landmarks.

2.2 Landmark Conceptualisation

In order to provide appropriate landmark-referenced route instructions, it is necessary to understand how landmarks can be used in spatial sentences to locate a trajector. On a geometric level, all landmarks can be considered areas when projected onto a top-down map. However, on a conceptual level, landmarks can be used in a point-like, line-like or area-like manner, depending on their spatial relationship with a route (Hansen et al., 2006).

One possible approach to determining a landmark's conceptual class is to make use of the landmark's geometric context, including its size relative to the route and the number of decision points with which it coincides. However, this approach may have little ecological validity, as people may not in fact conceptualise landmarks as point, line or area based purely on geometry, but also based on pragmatic considerations. For instance, it may be the case that people don't tend to conceptualise Flinders St Station as an area, even though it satisfies the geometric criteria.

2.3 Past Research on Landmark Interpretation

The only research we are aware of which has addressed this same topic of landmark interpretation is that of Tezuka and Tanaka (2005). In an investigation of the spatial use of landmarks in sentences, Tezuka and Tanaka (2005) modified existing web mining methods to include spatial context in order to obtain landmark information.

It is natural to question the appropriateness of web data for research purposes, because web data is inevitably noisy and search engines themselves can introduce certain idiosyncracies which can distort results (Kilgarriff and Grefenstette, 2003). However, the vast amount of data available can nevertheless give better results than more theoretically motivated techniques (Lapata and Keller, 2004). And importantly, the data that can be gleaned from the web does not mirror the view of a single person or a select group, but of the entire global community (or at least the best available representation of it).

3 Methodology

The prepositions and verbs which accompany a landmark in spatial sentences capture that landmark's implicit conceptualisation. We use this implicit conceptualisation, as represented on the web, to develop two automated classification schemes: a simple voting classifier and a neural network classifier. We compile a set of gold standard classifications in order to evaluate the performance of the classifiers.

3.1 Landmarks

A list of 58 landmarks was generated for Melbourne, Australia. The landmarks were chosen to be uniquely identifiable and recognisable by most inhabitants of Melbourne.

3.2 Gold Standard

We had annotators use a web interface to uniquely classify each landmark as either point-, line- or area-like. Each landmark's gold standard category was taken to be the category with the greatest number of annotator votes. Where the annotations were split equally between classes, the maximal geometric class was chosen, which is to say, line was chosen in preference to point, and area was chosen in preference to line. The rationale for this is that, for example, a point-like representation is always recoverable from a landmark nominally classified as an area, but not the other way around. Hence the classification which maintains both pieces of information, that this landmark may be treated as an area or a point, was assigned preferentially to the alternative, that this landmark may only be treated as a point.

Since landmark conceptualisations can depend on the mode of transport involved, annotators were instructed to consider themselves a cyclist who nevertheless behaves like a car by always staying on the street network. The intention was to elicit conceptualisations based on a modality which is intermediate between a car and a pedestrian. Annotators were also asked to indicate their confidence in each annotation.

3.3 Web Mining

We identified a set of prepositions and verbs as indicating a point-like, line-like or area-like representation. The number of documents on the web which were found to contain a particular landmark in point-like, line-like or area-like spatial sentences provided the raw data for our automated classification schemes. The web data thus obtained can be considered an implicit representation of a generalised cognitive model of the landmarks.

Prepositions

Landau and Jackendoff (1993) investigated the use of English spatial prepositions and the requirements they place on the geometric properties of reference objects. This analysis was projected onto the conceptual classes of point, line and area, to form a list of conceptually grouped spatial prepositions. Hence prepositions which require the reference object to be (or contain) a bounded enclosure, such as *inside*, were classified as denoting an area-like landmark; prepositions which require the reference to have an elongated principal axis, such as *along*, were classified as denoting a line-like landmark; and prepositions which place no geometric constraints on the reference object, such as *at*, were classified as denoting a point-like landmark.

The prepositions used were restricted to those which pertain to a horizontal planar geometry compatible with route directions; for example, prepositions which make use of a reference object's vertical axis such as *on top of* and *under* were ignored, as were prepositions denoting contact such as *against*. The preposition *out* was also excluded from the study as it is typically used in non-spatial contexts, and in spatial contexts the reference object is usually covert (eg *he took his wallet out*) (Tyler and Evans, 2003). Conversely, *out of* is frequently spatial and the reference object is overt, so this compound preposition was retained. The complete list of prepositions used in the study is given in Table 1.

Verbs

In addition to the list of prepositions, a list of verbs was created based on the verb classes of Levin (1993), restricted to verbs of inherently directed motion which can be used in a phrase immediately preceding a landmark, such as the verb *pass* in the phrase *pass the MCG*; in other words, the chosen verbs can be used in a way which parallels the use of spatial prepositions, as opposed to verbs such as

Point-like	Line-like	Area-like
across from	along	around
at	alongside	across
after		in
away from		inside (of)
before		into
behind		out of
beside		outside (of)
in front of		through
near		within
next to		without
opposite		
past		
to		
to the left of		
to the right of		
to the side of		
toward		

Table 1: Prepositions used in this research (based on Landau and Jackendoff (1993))

Point-like	Line-like	Area-like
hit	follow	cross
pass		enter
reach		leave

Table 2: Verbs used in this research

proceed, which specify a motion but require a preposition for clarification. This second type of verb is of no interest to the study as they tell us nothing about the conceptualisation of landmarks.

As with the prepositions, the verbs were grouped into the conceptual classes of point, line and area according to the requirements they place on reference objects, including *enter* for an area-like object, *follow* for a line-like object and *pass* for a point-like object. The complete list of verbs used in the study is given in Table 2.

Document Counts

Each of the prepositions and verbs was combined with each of the landmarks to create a cross-product of linguistic chunks, such as *at Queen Victoria Market*, *through Queen Victoria Market*, and so on. Alternative names and common misspellings of the landmark names were taken into account, such as *Flinders St Station*, *Flinders Street Station* and *Flinder's Street Station*. Additionally, three conjugations of each verb were used—present tense non-3rd person singular (eg *reach*), present tense 3rd person singular (eg *reaches*), and past tense (eg *reached*).

Each linguistic chunk was sent in turn to the Google search engine, which determined the approximate number of documents on the web containing that exact phrase. The counts were then summed over the conceptual categories in which each preposition and verb appeared. The result of this was a probabilistic categorisation of each landmark as point, line or area, according to its usage in spatial sentences on the web.

It is difficult to determine the context of sentences using a search engine. It is uncertain whether the documents found by Google use the searched-for linguistic chunks in a spatial context or in some other context. For this reason, each preposition and verb was assigned a weight based on the proportion of occurrences of that word in the Penn Treebank (Marcus et al., 1993) which are labelled with a spatial meaning. This weighting should give an approximation to the proportion of spatial usages of that word on the web.

Automated Classification

As a naive automated classification of the landmarks, the document counts were used to place each landmark in one of the three conceptual classes. Each landmark was placed in the class in which it was found to appear most frequently, based on the classes of the prepositions and verbs with which it appeared on the web. Hence landmarks which appeared more often with a point-like preposition or verb, such as *at* or *pass*, were placed in the point category; landmarks which appeared more often with a line-like preposition or verb, such as *follow*, were placed in the line category; and landmarks which appeared more often with an area-like preposition or verb, such as *around*, were placed in the area category.

As a more sophisticated classification scheme, we developed a supervised artificial neural network classifier. The neural network we developed consisted of a three node input layer, a two node hidden layer and a two node output layer, with learning taking place via the backpropagation algorithm. For each landmark, the percentage of web counts in each of the three conceptual classes was used as the initial activation value of the three nodes in the input layer. The activation of the output nodes was rounded to 1 or 0. The output node activations were used to indicate whether a landmark falls into the point, line or area category — 01 for point, 10 for line and 11 for area. An output of 00 was taken to indicate a failure to classify. The neural network was trained and tested using fourfold cross-validation, with the gold standard classification as the desired output in each case.

4 Results

Five experiments were conducted on the simple voting classifier and the neural network classifier. These experiments used increasingly sophisticated inputs and gold standard measures to try to improve the performance of the classifiers, as measured against the gold standard. The neural network classifier outperformed the voting classifier in all experiments but the final one.

Of the 58 Melbourne landmarks, 27 were classified as points by the majority of annotators, 2 as lines, and 29 as areas. These majority classifications were used as the gold standard. For these classifications, we calculated a kappa statistic of 0.528 (Carletta, 1996). This suggests that the annotation classification task itself was only moderately well-formed, and that the assumption that multiple annotators will classify landmarks in a similar manner does not necessarily hold true.

To determine whether the classifiers were performing at an acceptable level, we established a majority-class baseline: 29 of the 58 landmarks were areas, and hence the majority class classifier has an accuracy of 50%.

The maximum meaningful accuracy that can be achieved by a classifier is limited by the accuracy of the annotations themselves, creating an upper bound for classifier performance. The upper bound was calculated as the mean pairwise inter-annotator agreement, which was determined to be 74.4%.

	Accuracy (%)	E.R.R. (%)
Baseline	50.0	
Voting Classifier	63.8	56.6
Neural Net Classifier	70.0	82.0
Agreement	74.4	

Table 3: Results with simple web counts (Experiment 1)

4.1 Experiment 1

Experiment 1 involved using only the raw web count data as input into the classifiers. The accuracy and error rate reduction (E.R.R.) of the classifiers are given in Table 3.

The neural network classifier produced results slightly better than the simple voting classifier, but with 18 landmarks incorrectly classified by the neural network, there is still plently of room for improvement. The raw web count data used in this experiment was likely to be biased in favour of certain prepositions and verbs, because some of these words (such as *at* and *in*, which each occur in over 7 billion documents) are much more common than others (such as *beside*, which occurs in just over 50 million documents). This may result in the web counts being unfairly weighted towards one class or another, creating classifier bias.

The simple voting classifier showed a tendency towards point classifications over line or area classifications. The neural network classifier reversed the bias shown by the simple voting classifier, with the area class showing high recall but low precision, resulting in a low recall for the point class. Neither of the two line landmarks were classified correctly; in fact, none of the landmarks were classified as lines.

4.2 Experiment 2

To adjust for the potential bias in preposition and verb use, the web counts were normalised against the prior probabilities of the relevant preposition or verb, by calculating the ratio of the count of each linguistic chunk to the count of its preposition or verb in isolation. The accuracy and error rate reduction of the classifiers are given in Table 4.

Normalising the web counts by the prior probabilities of the prepositions and verbs did not improve the accuracy of the classifiers as expected. The sim-

	Accuracy (%)	E.R.R. (%)
Baseline	50.0	
Voting Classifier	55.2	21.3
Neural Net Classifier	70.0	82.0
Upper	74.4	

Table 4: Results with normalised web counts (Experiment 2)

ple voting classifier reduced in accuracy, while the accuracy of the neural net classifier remained unchanged.

4.3 Experiment 3

As explained in Section 3.2, the annotators who generated the gold standard were required to choose one of point, line or area for each landmark, even if they were unfamiliar with the landmark. Some of these annotators may have been forced to guess the appropriate class. As a result, these annotations may cause the gold standard to lack validity, which could be one of the barriers to classifier improvement.

In this experiment, a more sound gold standard was generated by weighting annotators' classifications by their familiarity with the landmark. The effect of this is that the judgement of an annotator who is very familiar with a landmark outweighs the judgement of an annotator who is less familiar. Experiments 1 and 2 were conducted again based on this new gold standard. These repeated experiments are dubbed Experiments $1'$ and $2'$ respectively. The results of each of the repeated experiments are shown in Table 5.

The simple voting classifier showed improvement using the weighted gold standard, with the accuracies under Experiments $1'$ and $2'$ each exceeding the accuracy of the equivalent experiment using the original gold standard. Experiment $1'$ showed the most improvement for the simple voting classifier, giving an accuracy of 67.2% (only one landmark shy of the 70% accuracy achieved by the neural network classifier in experiment 1).

While landmarks well-known to all are likely to produce consistently valid classifications, and landmarks poorly known to all are likely to produce consistently invalid classifications, regardless of whether a weighting scheme is used, it is the landmarks which are well-known to some and poorly known to others which should have gained the greatest benefit from annotations weighted by familiarity. However, the majority of such landmarks were already being classified correctly by the neural network in both Experiments 1 and 2, which explains why the neural network showed no improvement.

5 Discussion

Surprisingly, the naive conditions in Experiment 1 produced the best overall result, which was a 70% accuracy for the neural network classifier. Although the voting classifier and the neural network classifier produced similar levels of accuracy for many of the experiments, there was very little overlap in the landmarks that were correctly assigned by each classifier. Of the 40 landmarks correctly assigned by the neural network, 18 were incorrectly classified by the voting classifier. Conversely, of the 37 landmarks correctly assigned by the voting classifier, 15 were incorrectly assigned by the neural network. This indicates that the neural net is doing something more sophisticated than simply assigning each landmark to its maximum category.

A rather large subset of the landmarks was found to be consistently misclassified by the neural net, under various training conditions. For a number of these landmarks, the annotators showed strong disagreement and indicated that the landmark is ambiguous, suggesting that there is indeed an inherent ambiguity in the way these landmarks are conceptualised, both between annotators and on the web. Interestingly, all of the hospitals in the landmark list were consistently misclassified. A number of annotators expressed confusion with regard to these landmarks, as to whether the hospital itself or the surrounding gardens should be taken into account. As a result, annotations of the hospitals tended to be split between point and area.

However, some of the landmarks that were misclassified by the neural net were classified consistently by the annotators — for example, GPO was classified as a point by all of the Melbourne annotators. The ambiguity here presumably lies in the web counts, which were not able to detect the same conceptualisation generated by the annotators. One complication with using web counts is the fact

	Voting Classifier		Neural Network Classifier	
Experiment	Accuracy (%)	E.R.R. (%)	Accuracy (%)	E.R.R. (%)
$1'$	67.2	70.5	65.5	63.5
$2'$	58.6	35.2	65.5	63.5

Table 5: Results weighted according to landmark familiarity (Experiments $1'$ and $2'$)

that the data is global in scope, and with a simple abbreviation like GPO, there may well be interference from documents which do not refer to the Melbourne landmark, and in fact may not refer to a landmark or spatial object at all.

One of the underlying assumptions of the study was that all landmarks can be represented as falling into exactly one of the three conceptual classes — point, line or area. This may be an oversimplification. Some landmarks may in fact be more prototypical or ambiguous than others. Certainly, a number of the landmark annotations were split almost equally between point, line and area. It may be that annotators did not or could not take upon themselves the mentality of a cyclist as requested in the annotation instructions, and instead simply conceptualised the landmarks as they usually would, whether that entails a pedestrian or car modality, or some alternative such as a train or tram-like modality. It may also be the case that there are individual differences in the way people conceptualise certain types of landmarks, or indeed space in general, regardless of the modality involved. If this is true, then the low inter-annotator agreement may be a product of these individual differences and not merely an artifact of the experiment design.

In summary, we have proposed a method for classifying landmarks according to whether they are most point-like, line-like or area-like, for use in the generation of route descriptions. Our method relies crucially on analysis of what prepositions and verbs the landmarks co-occur with in web data. In a series of experiments, we showed that we are able to achieve accuracy levels nearing inter-annotator agreement levels for the task.

One simplification made during the course of this study was the treatment of parks and districts as being comparable entities (i.e. area-like landmarks). In fact, a distinction may be made between open areas such as districts, with which the preposition *through* may be used, and closed areas such as parks, for which *through* does not apply for car navigation (although obviously does apply for pedestrian navigation). We hope to take this into account in future work.

Acknowledgments

This research was supported by Australian Research Council DP grant no. DP0770931. The authors wish to thank Lars Kulik for his input into this research.

References

Gary L. Allen. 1997. From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. In *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science (COSIT 1997)*, pages 363–372.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Robert Dale, Sabine Geldof, and Jean-Philippe Prost. 2005. Using natural language generation in automatic route description. *Journal of Research and Practice in Information Technology*, 37(1):89–105.

Stefan Hansen, Kai-Florian Richter, and Alexander Klippel. 2006. Landmarks in OpenLS — a data structure for cognitive ergonomic route directions. In *Geographic Information Science — Fourth International Conference, GIScience 2006*, pages 128–144.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

Alexander Klippel. 2003. *Wayfinding Choremes: Conceptualizing Wayfinding and Route Direction Elements*. Ph.D. thesis, Universitt Bremen.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial cognition. *Behavioral and Brain Sciences*, 16:217–65.

Mirella Lapata and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of nlp tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 121–128.

Beth Levin. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. University of Chicago Press, Chicago.

Kirstin Lovelace, Mary Hegarty, and Daniel R. Montello. 1999. Elements of good route directions in familiar and unfamiliar environments. In *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science (COSIT 1999)*, pages 65–82, Stade, Germany.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–30.

Pierre-Emmanuel Michon and Michel Denis. 2001. When and why are visual landmarks used in giving directions? In *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science (COSIT 2001)*, pages 292–305, Morro Bay, USA.

Taro Tezuka and Katsumi Tanaka. 2005. Landmark extraction: A web mining approach. In *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science (COSIT 2005)*, pages 379–396, Ellicottville, USA.

Andrea Tyler and Vyvyan Evans. 2003. *Lexical meaning and experience: the semantics of English prepositions*. Cambridge University Press, Cambridge, U.K.