# Statistical Language Models For Topographic Data Recognition

Adam Winstanley, Bashir Salaik, Laura Keyes
Department of Computer Science
National University of Ireland, Maynooth
Maynooth, County Kildare, Ireland
adam.winstanley@may.ie

***Abstract*—The success of Statistical Language Models (SLMs) at improving the performance of Natural Language Processing (NLP) applications suggests their possible applicability to the area of automated map reading. This idea stems from the fact that there are similarities between natural language and cartographic language. We describe a method of using SLM to characterise the context of different classes of objects. We use these models to measure the frequency of each feature context. This can be used to help identify unclassified map features in combination with other methods (for example, based on an object's shape).**

## I. INTRODUCTION

To manipulate, analyse and retrieve topographic data in Geographical Information Systems (GIS), it is necessary to attach semantic information to the objects depicted. For many purposes, these meanings are standardized in the form of a set of standard categories, sometimes grouped or classified into successively higher levels of abstraction (super classes). Thus we may talk of a buildings class consisting of industrial, domestic and commercial sub-classes (figure 1). Each of the subclasses can usually be further subdivided according to the needs of the application.

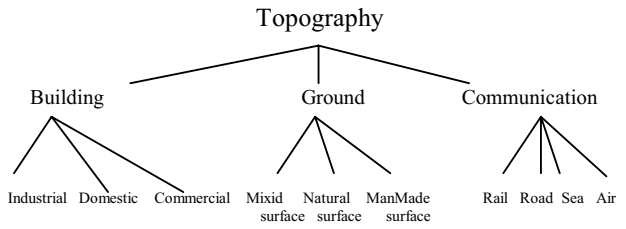


Figure 1. Hierarchical classification system.

In most cases, this data classification process is performed manually or semi-automatically during the secondary data acquisition stage, which may involve manual digitizing or an automated approach (for example though detection of line thickness) when raster scanning paper maps. Even when automated, the digital map may have classification errors from the data capture stage, and so an alternative method of classifying data (or checking the existing categorisation) is desirable.

In this paper we describe an approach of using Statistical Language Models (SLM) to solve this problem. These models have had a great success at improving Natural Language Processing (NLP) applications. They work by building a statistical model of word associations through analysis of a standard corpus. This model can then be used to check, for example, the output from a speech processing system for invalid or unlikely phrases.

There is a linear (one dimensional) structure of natural language utterances and an underlying grammatical structure of the adjacent language units (for example, words and sentences). The use of these models in structuring topographic data is motivated by the analogy between natural language and cartographic language [1].

- Both types of data consist of discrete objects (words, map-feature/objects).
- Objects have a physical form (spelling, shape).
- Objects have a semantic component (meaning, feature class).
- Objects are formed into larger components (sentences, regions/map-sheets and so on).

In topographic data, there is no strict *grammatical* structure between the objects depicted. However a quasi-grammatical pattern does exist (for example, house-garden-road sequences in sub-urban areas). This suggests that the language modeling approach may have some validity. However, unlike natural language, the topographic "sentences" have no inherent direction be followed.

## II. STATISTICAL LANGUAGE MODELS

SLMs were first applied by Andrea Markov at the beginning of the 20th century to model letter sequences in works of Russian literature [8]. Later, SLMs were developed as general natural language processing tools and language models were first applied to automatic speech recognition in the 1970s. SLMs were also applied to many other areas of natural language processing, for instance machine translation [3] and part-of-speech tagging [5]. In geographical analysis, Markov models have also been applied to temporal and spatial data, for example, characterising migration patterns, land use and land cover change [2, 4].

Sponsored by the Enterprise Ireland, British Council, and Ordnance Survey (Southampton), UK

Statistical language modeling is an attempt to capture regularities in natural language for the purpose of improving the performance of various processing applications. In general, SLMs consist of estimating the probability distribution of linguistic units such as words, sentences and whole documents and using this to predict the next unit in a sequence. The place of an SLM in a possible system architecture (to improve speech recognition) is shown in figure 2.

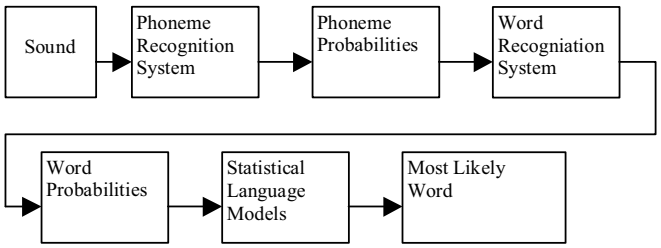


Figure 2. Typical speech recognizer

SLMs employ statistical estimation techniques using language-training data in the form of corpora of text. Due to the categorical nature of language, and the large vocabularies people naturally use, statistical techniques must estimate a large number of parameters, and consequently depend critically on the availability of large corpora.

A statistical language model is a probability distribution over a sequence of words. A language model is represented as a conditional probability of the next word in a sequence ($w_i$) given the previous words ($h_{i-1}$), that is

$$P(w_i \mid h_{i-1}), \text{ where } h_{i-1} = w_1, w_2, .....w_{i-1} \qquad \text{(equation 1)}$$

Different *n-gram* models can be constructed depending on the length *n* of the word sequences used, for example, uni-gram (i=1), bi-gram (i=2) and tri-gram (i=3). These models can be combined using linear interpolation, for example,

$$p(w \mid h_i) = \lambda_1 p_1(w \mid h_i) + \lambda_2 p_2(w \mid h_i) + \lambda_3 p_3(w \mid h_i)$$

$$\text{where} \quad \lambda_1 + \lambda_2 + \lambda_3 = 1, \; h_i \geq 0 \qquad \text{(equation 2)}$$

The purpose of these models is to assign high probabilities to likely word sequences and low probabilities to unlikely ones.

There are many statistical language models, but we only focus on the most powerful models (n-gram) because:

- These models have the advantage of being able to cover a much a larger language than would normally be derived from a corpus.
- In contrast to grammatical language models, n-gram models rely on the likelihood of sequence words such as word pairs, (bi-gram) or word triples (tri-gram) therefore they are less restrictive.
- Open vocabulary applications are easily supported with n-gram models.
- The use of n-gram models has a long successful history in the research community and is now used in commercial systems.

## III. SLMs Applied to Topographic Data

A possible architecture for a topographic object recognition system, analogous to the speech recognition system described earlier, is given in figure 3. The image is vectorised, cleaned and topologically corrected to form closed polygons. A recognition system produces probabilities for candidate classes of each object based, in this case, on their shape [7]. The SLM, built from analysis of another data-set, uses the probabilities to construct "phrases" of objects and use the n-gram model built from a corpus to select the most the most likely phrases.

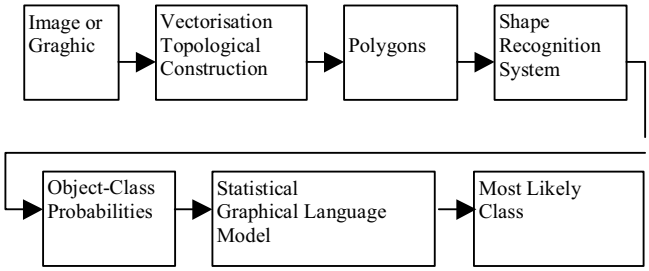


Figure 3. A possible topographic object recognition system

Given the similarities identified between language and cartography, it seems reasonable that an SLM may improve the performance of topographic object classifiers in the same way as they do for language processing applications. One major difference is that, whereas language is essentially a one-dimensional sequence of symbols, maps are inherently two-dimensional. Therefore it is necessary to extract one-dimensional sequences from the topographic data. One way of doing this is to use the adjacencies between objects (polygons) on the map. For example, in figure 3 the central polygon (H) representing a house is surrounded by a garden (G), field (F) and road (R). There are four different phrases present in this pattern excluding those that double back on themselves. Two of these terminate at H.

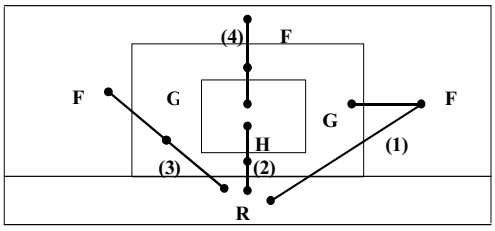


Figure 4. Extracting "phrases" from topographic data.

*Uni-gram Model*

The uni-gram model is used to calculate the frequency of each class on the map, and so is only a crude prediction tool. By applying this model, we can estimate the probability of a certain object belongs to a particular feature class irrespective of context.

*Bi-gram Model*

To construct the bi-gram probability from the data-set, we count the number of occurrences of object pairs $w_{n-1}, w_n$ and divide that by the number of occurrences of the class $(w_{n-1})$.

*Tri-gram Model*

To construct this model, for each object in context, we count the number of times the triple $(w_{n-2}, w_{n-1}, w_n)$ is observed and divide this by the number of times the pair $(w_{n-2}, w_{n-1})$ occurs.

*The Fusion Model*

Unlike natural language, on the map each object forms part of several phrases, each of which produces a candidate class with an associated probability. To produce a single final classification decision, data fusion techniques proposed by Kittler [6] were implemented to combine these opinions, namely the *min*, *max*, *sum*, *product*, *median* and *majority vote* strategies. While the performance of each classifier was variable, combining them in one scheme should yield the best performance.

## IV. EXPERIMENTAL RESULTS

Three n-gram models (i=1,2,3) were constructed based on the analysis of an Ordnance Survey 1:1250 data set (Master Map) representing part of the Basingstoke area in Great Britain. The data set contained 67,805 objects in 13 classes (table 1). The class of each object was checked using the n-gram models. The results from the models were then combined using the product fusion scheme described by Kittler [6]. The class of each object was taken to be the maximum likelihood class for that object output by the model. This final reclassification was compared to the original object description to evaluate the predictive effectiveness of the models.

TABLE I. PERFORMANCE OF BIGRAM SLM ON SAMPLE DATA SET.

| OS Code | *Description* | *Number* | *correctly classified* | |
|---|---|---|---|---|
| 10021 | Building | 29752 | 20030 | 67.3% |
| 10053 | Ground (mixed surface) | 25288 | 18774 | 74.2% |
| 10054 | Ground (natural surface) | 378 | 0 | 0% |
| 10056 | Ground(manmade surface | 5029 | 1353 | 26.9% |
| 10062 | Glass-roofed building | 25 | 0 | 0% |
| 10089 | Inland Water | 18 | 0 | 0% |
| 10111 | Vegetation cover | 461 | 0 | 0% |
| 10123 | Paths | 598 | 0 | 0% |
| 10167 | Rail | 14 | 2 | 14.2% |
| 10172 | Road and Tracks | 1842 | 0 | 0% |
| 10183 | Road Side | 3222 | 779 | 24.1% |
| 10185 | Generalmanmade Structure | 42 | 17 | 40.4% |
| 10217 | Unclassified | 18 | 0 | 0% |

Table 1 shows the performance of the bi-gram model, showing the number of polygons and the percentage correctly classified for each class. The results were reliable for feature classes that had high populations but very poor for other classes with low populations. In particular, buildings and gardens were classified the most reliably. This performance was increased by combining classes from the same super-class (table 2).

TABLE II. CLASSIFICATION PERFORMANCE USING TWO SUPER CLASSES.

| **Description** | ***Number*** | ***correctly classified*** | ***Percentage %*** |
|---|---|---|---|
| Building | 29777 | 20030 | 67.3% |
| Ground | 30693 | 20127 | 65.5% |

## V. CONCLUSION

The preliminary results presented here suggest that these statistical models could be a helpful tool to structure common classes of object through their context by identifying unclassified and/or miss-classified features. Compared to corpora used in natural language models, the data set used here is small and so the results obtained are only indicative for the most common classes. Further work needs to be done with much larger dataset. However, even we if we were to use a larger data set, sequences involving rare classes will be scarce and this technique is unlikely to classify them.

It is envisaged therefore that statistical language models will in practice be used in combination with other methods base on shape and context and will be used to improve or modify the confidence in the classification obtained from those models. More investigation needs to be done with larger corpora of different characteristics (varying region types, scale and so on), different fusion models and in combination with other object recognition techniques.


## ACKNOWLEDGMENT

Our sincere thanks go to Ordnance Survey (Great Britain) for providing the data for this project. This work was partly supported by an Enterprise Ireland/British Council Research Visits Scheme Grant (BC/2002/015).


## REFERENCES


[1] J.H. Andrews, "Maps and language, a metaphor extended", Cartographic Journal, Vol. 27, pp. 1–19, 1990.

[2] E.J.Bell "Markov analysis of land use change - an application of stochastic process to remotely sensed data", Journal of Socio-economic Planning Sciences, Vol. 8. pp.68–73. 1974.

[3] P.F.Brown, J.Coke, S.A.D.Pietra, V.J.D.Pietra, F.Jelinek, J.D.Lafferty, R.L.Mercer and P.S.Rossin, "A statistical approach to machine translation" Computational Linguistics, Vol. 16(2), pp. 79–85, 1990.

[4] L.Collins An introduction to Markov Chain Analysis, Concepts and Techniques in Modern Geography 1, Geo Abstracts Ltd., Norwich 1974.

[5] D.,Cutting, J.,Pedersen, and P.Sibun "A practical part-of-speech tagger," Third Conference on Applied Natural Language Processing (ANLP-92), pp.133-140, 1992

[6] J.Kittler ,M.Hatef., R.P.W.Duin and J.Matas, "On combing classifiers" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20(3), pp. 226-239, 1998.

[7] L.Keyes and A.C.Winstanley, "Fourier descriptors as a general classification tool for topographic shapes", Proceedings of the Irish Machine Vision and Image Processing Conference, pp.193-203, Dublin City University, 1999.

[8] C.D.Manning and H.Schutz "Foundations of Statistical Natural Language Processing", MIT Press, Cambridge, 1998.