Chapter 7

Conclusions

Probabilistic language modelling is central to a wide variety of applications, from data compression to speech recognition. The field has a long history and much progress has been made over the years. However, this progress has been in many cases due to heuristic approaches which while improving predictive performance, do not necessarily increase our understanding of the underlying processes.

One of the key contributions of this thesis has been to demonstrate the relationship between an important existing approach, generalised PPM-A, and the hierarchical Dirichlet language model, which is a simple Bayesian model based on a clear theoretical framework. Although exact inference in the latter is not possible on a realistic time scale, this thesis has shown that generalised PPM-A is an excellent approximation to this model under realistic conditions. Many applications of language modelling, for example speech recognition, require real-time computation, whereas others such as information retrieval must process many thousands of documents within a time-frame which is acceptable to the user. These applications therefore present considerable restrictions on the available computational resources, making low cost approximations such as generalised PPM-A invaluable.

Using generalised PPM-A as a starting point, it was demonstrated that basic symbol-level language modelling can be extended to make use of higher level structure present in the text, namely the knowledge that the text stream consists of a sequence of separate word and non-word characters. This information permits the construction of a model which is able to make use of a word list, even if relative frequencies of the terms are not known.

This approach was shown to be beneficial when either no training text is available or the quantity of training text is very small, a result which in itself is potentially beneficial. However, the model presented in Chapter 3 did not perform as well as a simple symbol-level language model when large quantities of training text were available. To some extent this result is disappointing, as one might have expected a more sophisticated model with access to the same training data to be able to perform at least as well. This area can therefore be identified as one where there is potential for further investigation, with the aim of improving on the results presented here.

Chapter 4 investigated the application of hierarchical Dirichlet models to information retrieval. This approach was shown to perform favourably when compared to other models, giving an improvement in performance in at least some cases. However, perhaps more importantly, it was shown that the resulting method naturally includes document frequency information as a way to identify the most discriminative terms in the query. This information has long been used in information retrieval, but until now has either been missing or introduced in an *ad hoc* way in those approaches which have made use of language models. The natural inclusion of this information further demonstrates that much progress can be made using simple, but principled, probabilistic methods.

There is much scope for further work in the application of language modelling to information retrieval. Chapter 4 demonstrated one direction for further investigation by extending the hierarchical model to model passages within the documents. The resulting method was able to consider not only terms occurring within each passage, but also within the context of the document of which the passage is a part. This approach gave promising results when compared to the simpler method of treating passages as separate documents. However, it is likely that further progress can be made, for example by investigating alternate definitions of passages or looking at the possibility of a continuously varying measure of relevance throughout the document. Further examples of areas which could be considered are the use of 'topic' information in retrieval, and the use of information on word correlations within documents. In both of these areas the development of simple approximations is likely to be important.

Having looked at probabilistic models in the context information retrieval, Chapter 5 looked at the Dasher interface as a way of extending the use of probabilistic information to the process of entering search queries and to displaying the results. A prototype implementation of an interactive search tool for locating substrings in a text document showed that this approach is viable. A further investigation of this method as a means of locating contact information indicated that, at least in terms of a theoretical analysis, substantial improvements to the efficiency of the search process are possible. To complete this investigation, it is now necessary to evaluate the performance of these applications in trials on real users.

The final chapter in this thesis considered richer documents than those containing just plain text. In particular, the chapter focused on the task of interpreting hand-drawn electronic ink diagrams, which are entered using a pen-based interface. This model was shown to perform well on example data, and in particular it was shown that the use of partitioning information was able to improve labelling performance over models which considered labelling only.

There are a number of further applications of the model developed for diagram analysis, for example in segmenting bitmap images or video. It is likely that these applications will involve more complex inference tasks, so development of approximate methods will be required. Possible approaches to this problem have already been considered in Chapter 6.