LM vs PM: where's the relevance?

Karen Sparck Jones
Computer Laboratory, University of Cambridge
Stephen Robertson
Microsoft Research, Cambridge

ksj@cl.cam.ac.uk, ser@microsoft.com

The Language Modelling (LM - see refs 'LM') approach to the relation between a document and a request is to ask: how probable is it that this document generated the request? More strictly the question is: how probable is it that the document, as represented by its index description, generated the request as represented by its indexing description? There is no explicit appearance here of what has hitherto been regarded as the key retrieval notion, namely relevance. However the presumption is that if it is highly probable that the document generated the request, then the document's content is relevant to the information need underlying the user's request. A good match on index keys implies relevance, though relevance is never mentioned in the model.

The Probabilistic Modelling (PM) approach asks a quite different question. That is, hijacking "Probabilistic Modelling" to refer to models in the Maron-Robertson tradition instantiated in Okapi (refs 'PM'), the PM approach asks: how probable is it that the document is relevant to the request? More properly the question is: how probable is it that the document content is relevant to the user's need? Then since both content and need are unobservables, it is necessary to provide an argument for taking index descriptions as levers in deriving the operational equivalent of probability of relevance. The PM approach thus leads to the same position as the LM one: a good match on index keys implies relevance. However, in contrast to the LM approach, relevance figures explicitly in the PM model.

But does this difference between the formal models matter when their operational interpretations, and hence manifestation for the user, are the same? The LM approach might seem to have the advantage that it does not deal directly in unobservables, since even if it allows for different forms of index description these are still derived from observable initial documents and requests. But of course the LM is being applied to retrieval, which deals in content and need, so this surface distinction between LM and PM is somewhat spurious and does not really matter. There is however a much more important difference between them, which matters much more.

It is a fact of retrieval life that there may be more than one relevant document for a request; indeed there are normally several. It is also a fact of retrieval system life that relevance feedback, exploiting knowledge of some relevant documents to help with retrieving others, is useful.

Formally, LM takes each document and, using its individual model in conjunction/comparison with a generic model (the file model), asks how likely it is that this document generated the request. The outcome is different values for documents, conveniently leading to a ranking.

For practical experiments this output is treated just like that for any other retrieval methods inducing a ranking, i.e. as delivering multiple relevant documents if these exist, and as having a performance measured using rank-based methods. However the principle underlying the model (at least as typically presented) is that it is identifying *the* document that generated *the* request. This document ought to be the one with the highest value, but since the base for estimating value is not perfect, the generating (i.e. relevant) document could be another one. However once this relevant document is recovered, retrieval stops.

The way the LM approach is applied in practice, that is, subtly changes its meaning. As long as documents are really different, they should generate different requests. The actual request may not be the best they could generate (indeed, will obviously not be), but that's tough and incidental. Similarly if the document index descriptions used for generation are impoverished, e.g. consist of two or three subject headings as opposed to full text, several documents could be deemed equally likely to have generated a given request. But this again simply implies that the base for determining the real relative status of different document vis-a-vis a request is weak, not that there are no real differences of status.

Thus if we now consider the situation where we recognise the empirical fact that there may be several documents relevant to the user's information need, and where we also have some particular expression of that need as a request, what does this imply for the LM approach to retrieval?

We will suppose first, simply but unrealistically, that there is no feedback option. Then the LM approach most obviously works as follows. We determine how probable it is that each document generated the request. Then since each relevant document has to be viewed as independent of every other and, further, in assuming that this document generated the request we have formally to treat all the other relevant documents as non-relevant, we notionally conduct a series of independent searches for each relevant document alone. The logic is the same as if we actually had several needs each with their own request, which just incidentally happen to be worded in the same way. In practice, we might economise and emerge with the same single ranking as in conventional tests, but this is vulgar practice, not proper principle.

On the other hand, if we try to incorporate the reality that there may be several documents relevant to the need into the model, this formally implies not that we look for the single document that generated the request, but that we look for the set of documents that generated the request. That is, the relevant document set is that set of documents that most probably generated the request. But of course this way, madness lies. Moreover, given that the information available to determine the set is only the initial request, the chance that the most probable set will coincide with the relevant set is rather low.

Thus there really are good and sufficient reasons for asking how relevance feedback fits the LM approach. As the LM approach, like the PM one, benefits from having more information for estimating probabilities, it is rational to ask how knowing about some relevant documents can improve estimation, and hence retrieval, for others.

Suppose, therefore, that we do our first search and find our first relevant document, i.e. the document that generated our initial request. If we modify the request, say by adding terms taken from the document, the LM approach strictly requires that we now treat this as a new request and find the document that most probably generated it. But of course since the new request has been explicitly built from a document - in the limiting case we might just have adopted that document's description as the new request - we have to withhold that document in the search or we will simply go round in circles. We thus change the file (if only infinitesimally) as well as the request.

It is clearly possible to continue on this one by one basis. But it has to be accepted that, without any reference to the idea that there is an underlying need and that the process is intended to improve the expression of that need so as to increase the chance of identifying all the documents relevant to it, the LM approach implies that we are dealing with a succession of needs and corresponding requests. The real life analogue is routing rather than adhoc retrieval. Moreover, the one-by-one model of retrieval that is entrenched in the LM approach naturally implies a step-wise modification (i.e. replacement) of the last version of the request; and this does not necessarily lead to the same eventual request to retrieve the last relevant document as may be obtained when requests are modified (or replaced) using the information supplied by a set of relevant documents taken together.

Thus overall the LM approach, at the theoretical level, implies a kind of minimalism in relation to sets of relevant documents, i.e. with respect to the information they supply about one another and hence can jointly supply to leverage a form of request that best retrieves them all.

All this is not to imply that the LM approach, as applied in practice for instance in TREC, does not work well. The aim here is to be clear about what the LM as a theoretical basis for document retrieval involves, and hence about the liberties that practical applications take with the pure theory. Thus the claim is that the LM approach, when investigated, in fact relies on unobservables just as much as the PM approach does; but it is at the same time less comprehensive than the PM one, and hence offers a less satisfactory account than the PM approach does of significant elements of the retrieval situation.

References

LM:

Berger, A. and Lafferty, J. 'Information retrieval as statistical translation', *SIGIR-99, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, 222-229.

Berger, A. and Lafferty, J. 'The Weaver system for document retrieval', *Proceedings, the Eighth Text REtrieval Conference (TREC-8)*, Ed. E.M. Voorhees and D.K. Harman, Special Publication 500-246, National Institute of Standards and Technology, Gaithersburg MD, 2000, 163-173.

Hiemstra, D. 'A linguistically motivated probabilistic model for information retrieval', *ECDL: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, 1998, 569-584.

Hiemstra, D. *Using language models for information retrieval*, PhD Thesis, University of Twente, 2001.

Hiemstra, D. and Kraaij, W. 'Twenty-One at TREC-7: ad-hoc and cross-language track', *Proceedings, the Seventh Text REtrieval Conference (TREC-7)*, Ed. E.M. Voorhees and D.K. Harman, Special Publication 500-242, National Institute of Standards and Technology, Gaithersburg MD, 1999, 227-238,

Kraaij, W., Pohlmann, R. and Hiemstra, D. 'Twenty-One at TREC-8: using language technology for information retrieval', *Proceedings, the Eighth Text REtrieval Conference (TREC-8)*, Ed. E.M. Voorhees and D.K. Harman, Special Publication 500-246, National Institute of Standards and Technology, Gaithersburg MD, 2000, 285-299 (see paper Appendix).

Miller, D., Leek, T. and Schwartz, 'BBN at TREC-7: using hidden Markov models for information retrieval', *Proceedings, the Seventh Text REtrieval Conference (TREC-7)*, Ed. E.M. Voorhees and D.K. Harman, Special Publication 500-242, National Institute of Standards and Technology, Gaithersburg MD, 1999, 133-142.

Papineni, K. 'Why inverse document frequency?' *NAACL-01, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2001.

Ponte, J.M. , *A language modelling approach to information retrieval*, PhD Thesis, University of Massachusetts at Amherst, 1998.

Ponte, J.M. 'Language models for relevance feedback', in *Advances in information retrieval*, Ed. W.B. Croft, Dordrecht: Kluwer, 2000, 73-95.

Ponte, J.M. and Croft, W.B. 'A language modelling approach to information retrieval', *SIGIR-98, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, 275-281.

PM:

Sparck Jones, K., Walker, S. and Robertson, S.E. "A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2", *Information Processing and Management*, 36 (6), 2000, 779-808 and 809-840.