THE STATIONARY CHARACTER OF PROBABILISTIC LANGUAGE MODELS FOR TEXT RETRIEVAL

Eduard Hoenkamp
Nijmegen Institute for Cognition and Information
hoenkamp@acm.org

Abstract

Statistical language modeling is gaining recognition as a viable alternative to the vector space model. In the latter, documents are points in a metric space spanned by the terms, while in the former they are samples from a probability distribution over the terms.

The language models proposed to date are overly general, however. They overlook the property that texts are samples from a natural language corpus. In contrast this paper capitalizes on this property to derive an interesting characteristic of language models for texts. First it shows that the general language model can be simplified to a Markov model. Next it shows that for natural language modeling the Markov chain ergodic.

The theoretical result is that the joint distribution for language models can be easily obtained. For IR applications this means that language modeling systems need fewer ad hoc assumption to make them practicable.

Keywords: Language models, ergodic process, natural language.

1. INTRODUCTION

Recent years have shown a growing interest in probabilistic language models for IR. Besides text retrieval in general, applications can be found in areas such as topic detection (Beeferman, Berger, & Lafferty, 1999), query disambiguation (Allan & Raghavan, 2002), and cross-language retrieval (Berger & Lafferty, 1999).

In probabilistic language modeling, a document is seen as a sample from a term sequence generated according to some probability distribution (Ponte & Croft, 1998). This way, documents can be characterized by their corresponding distributions. These distributions would in turn assign different probabilities to a sequence of query terms. For a given a query, then, the documents are ranked according to the probabilities their distributions would assign. Hence, a language model for IR needs to describe (1) the document model, i.e. distribution over terms, and (2) the query model, i.e. how a probability is assigned given the query terms. A variety of proposals have been published regarding the choice for (1) and (2). A recurring problem, however, is to find a satisfactory definition for the joint distribution over the query terms. The paper proposes a solution to this problem that is theoretically plausible, and easy to apply in practice.

2. the language model as a markov chain

2.1 The adequacy of lower order dependencies

In the language modeling approach one needs to estimate the probability of a query Q given a document D, that is $Pr(Q|D)$. Given a set of documents D_k, their ranking would correspond to the values $Pr(Q|D_k)$. So for Q as the term sequence $q_1q_2...q_n$, the task is to find the joint probability $Pr(q_1, q_2, ..., q_n|D)$. If the terms were independent, $Pr(Q|D)$ would simlpy be $\prod_i Pr(q_i|D)$. In general, the occurrence of a term depends on its context, i.e. on the occurrence of previous terms. So, to find the actual $Pr(Q|D)$ authors have suggested to use bigrams, tri-grams, and higher orders using the Bayes' chain-rule (Song & Croft, 1999; Bruza & Song, 2003). The issues with the approach are well-known: in principle we have the combinatorics of dependencies needed to apply the chain-rule, and in practice we have the problem of short documents (i.e. small samples) and missing term probabilities.
Interstingly, there are good reasons why lower order dependencies suffice, empirical as well as theoretical:

- In practice, bigrams give a reasonable improvement over unigrams (Lafferty & Zhai, 2001). In addition, (Song & Croft, 1999) shows that an interpolation of unigram and bigram models performs well.

- Many cognitive phenomena can be understood sufficiently in terms of word-pairs. Pertinent examples can be found e.g. in the research on memory (Shiffrin & Steyvers, 1998) and on the 'semantic space' (Burgess, Livesay, & Lund, 1998).

Note that the practical reason could just show the adequacy of language models per se. The evidence from cognitive studies, however, supports the use of lower order dependencies in modeling documents produced by humans, i.e. formulated in natural language. In models for natural language, therefore, we need only consider simple priors. Looking at the process that the language model generates, let us identify its *state* with the term (word) generated. As only unigrams and bigrams need to be considered, the process has the Markov property, i.e. the probability of a state depends only on the previous state.

In summary, the language model can be represented as a Markov chain, with terms (words) as states, and the language samples as samples from the Markov chain[1].

[1] New words will enter a natural language, but the set will remain countable. A Markov process over countable states is called a Markov *chain*.

2.2 The document as an ergodic process

From the observation that the documents are samples of a natural language can we not only derive the Markov representation, but also an important property of the Markov process itself. To see this, note two important observations that seem so obvious that they are easily overlooked:

- Words in a natural language corpus can be separated by any number of intermediate words. (Think of adding an extra adjective before a noun.) This means there cannot be any cycles in the process. As we identified words with the states of the process, this means the Markov chain is *aperiodic*.
- You can always get from one word to another by continuing to produce text (words can never be used up). Consequently, the Markov chain is *irreducible*.

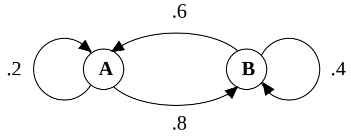
A Markov process that is both aperiodic and irreducible is called *ergodic*. An ergodic process has the property that in the long run it reaches a stationary distribution, irrespective of the initial state. **Box 1** below illustrates how of the distribution converges for an imaginary document from a corpus containing just two words.

Box 1

The underlying theory in a simple pedagogical example. Assume a language of just the words A and B, and a document with HAL matrix H (explained below). H is identified with the transition matrix of a Markov chain.

	A	B
A	.2	.6
B	.8	.4

$\Rightarrow$



For initial state s_0 (e.g. $\langle 1, 0\rangle$ if started with word A), the probability for the next state is given by $s_1 = s_0 * H$, where

$$H = \begin{pmatrix} .2 & .6 \\ .8 & .4 \end{pmatrix}$$

followed by $s_2 = s_1 * H = s_0 * H^2, ..., s_n = s_0 * H^n$ with

$$H^n = \frac{1}{.8 + .6} \begin{pmatrix} .6 & .6 \\ .8 & .8 \end{pmatrix} + \frac{-0.4^n}{.8 + .6} \begin{pmatrix} .8 & -.6 \\ -.8 & .6 \end{pmatrix} \tag{1}$$

which converges to:

$$\lim_{n \to \infty} H^n = \begin{pmatrix} .4286 & .4286 \\ .5714 & .5714 \end{pmatrix} \tag{2}$$

so the Markov chain becomes stationary with $P(A) = .4286$ and $P(B) = .5714$, independent of the initial state.

The use of Markov chains to find a joint probability is in itself not new. In domains with many dependencies, one often uses Metropolis-Hastings sampling (one of several MCMC (Markov Chain Monte Carlo) techniques): To find a joint probability, long chains of events are produced using Monte Carlo simulation. Next, the joint distribution is found by sampling from the chain. The method is illustrated in figure 1 for the priors given in **Box 1**. A difficulty inherent in the method is to establish whether the chain has converged to a point where sampling gives reliable probability estimates (Brémaud, 1999). In our experiments with the BM25 data (Robertson, Walker, Hancock-Beaulieu, Gull, & Lau, 1992) this method would be prohibitive: each and every query of the hundreds we tested would require millions of state changes besides the problem of knowing where to start sampling. Our method obviates the simulation of state changes and subsequent sample steps, and directly computes the joint distribution from the conditionals.

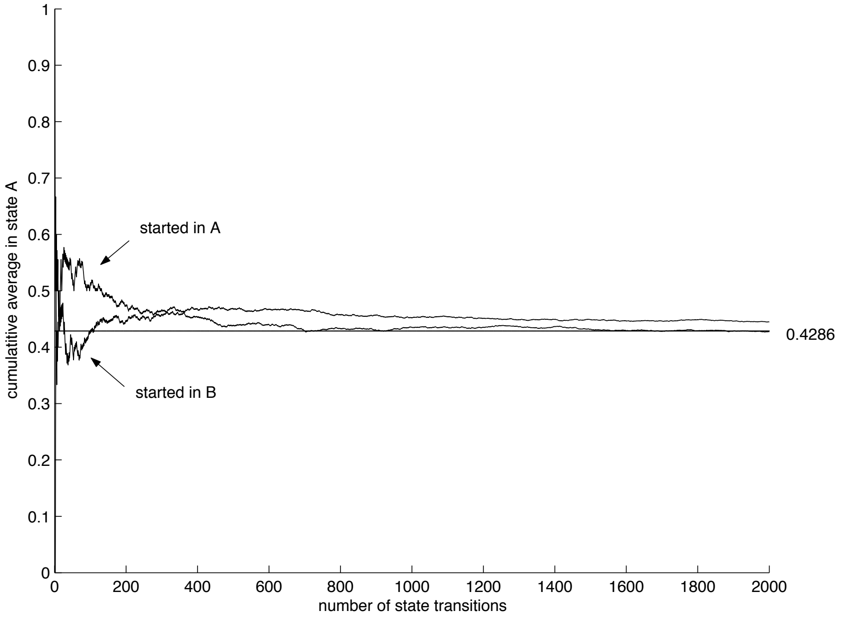


Figure 1: One way to find a joint probability is by Monte-Carlo simulation, as illustrated here using the Markov chain from **Box 1**. After a few thousand state changes, called the 'burn-in', the chain becomes stationary and it can be sampled to find the distribution. For a real corpus the burn-in takes millions of state changes, and special techniques are required to establish convergence. The method presented in this paper obviates such a simulation.

To recapitulate: (1) the states of the Markov chain represent the words of the intended corpus, (2) the transition probabilites represent the conditional probabilities of the bigrams, (3) the process has a stationary distribution. This distribution is easy to compute. It is also easy to characterize: it is the left eigenvector associated with eigenvalue 1 of the transition matrix. We propose to use this simple model of a stationary distribution in applications where others have approximated Bayes' chain rule. Those approximations rely on simplifying assumptions about the conditionals, and also on specialized techniques such as smoothing or interpolation.

Finally, note that the Markov chain defines one single, unstratified event space. Hence the method does not suffer from the kind of inconsistencies that Robertson(2002) observed in several language models.

2.3 Computing conditionals (bigrams)

Even if one agrees that the method just laid out is promising, there remains the question: where do the conditionals come from?

In our experiments, however, the conditional probabilities are derived from the 'Hyperspace Analog to Language' (HAL) representation for a corpus (Burgess et al., 1998). We prefer this method as it is rooted in early cognitive studies by Osgood et al. (1957) that measured how words are related in meaning, the so called 'semantic differential'. The HAL representation is computed by sliding a window over the documents and assigning weights to word pairs, inversely to the distance from each word to every other word in the window. In our experiments we normalized the weights to achieve a probability distribution. The resulting matrix is used as the Markov probability matrix.

As we have argued elsewhere (Hoenkamp, 2003), search may occur at other resolution levels than the document. At a finer level, searchers are more interested in a relevant passage then in the document itself. At a courser level, they may want a collection of documents handy when writing an article. Both levels may be modeled with a probability distribution, and both would generate an ergodic process.

We conducted an experiment on relevance feedback using the Markov chain approach. Here we have a rather course resolution level: the relevant set as a whole has a distribution that sets it apart from the non-relevant part of the corpus. We applied our method to Robertson's (1992) query expansion approach in TREC-3, improving significantly on precision and recall. Further details about the experiments are beyond the scope and aim of this paper[2].

[2] The results of the experiments will be published elsewhere, as this presentation aimed for a self-contained theoretical paper for the Formal/Mathematical Methods Workshop.

3. CONCLUSION

We believe that current language models are overly general because they do not incorporate properties of natural language, the very fabric of the documents they portend to model. This makes computing a joint distribution given the priors a recurring and difficult issue.

By making these properties explicit, we could derive a simple method to compute the joint distribution under conditions that are theoretically plausible, and easy to verify in practice.

The method represents the conditional probabilities as the transition probability matrix of a Markov chain, and we showed that this chain is ergodic. As a result, the eigenvector of the matrix, associated with eigenvalue 1 represents the joint distribution.

We hypothesize that the method can be used at different levels of search resolution, and since it is so simple it merits application in other areas where language models have been used.

ACKNOWLEDGEMENTS
I would like to thank Peter Bruza and Dawei Song for discussing the content of this paper during my stay at DSTC Brisbane, and Dawei in particular for running the experiments to corroborate the theoretical findings reported here.

References

Allan, J., & Raghavan, H. (2002). Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of SIGIR-2002* (pp. 307–314.

Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical models for text segmentation. *Mach. Learn.*, *34*(1-3), 177–210.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of SIGIR-1999* (pp. 222–229.

Brémaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*: New York: Springer.

Bruza, P., & Song, D. (2003). A comparison of various approaches for using probabilistic dependencies in language modeling. In *Proceedings of SIGIR-2003.*

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*, 211 – 257.

Hoenkamp, E. (2003). Unitary operators on the document space. *Journal of the American Society for Information Science and Technology*, *54*(4), 314–320.

Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for IR. In *Proceedings of SIGIR-2001* (pp. 111–119.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*: Urbana: University of Illinois Press.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR-1998* (pp. 275–281.

Robertson, S. (2002). On bayesian models and event spces in information retrieval. In *Proceedings of the Mathematical/Formal Methods in Information Retrieval at SIGIR-02*.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, A., & Lau, M. (1992). Okapi at TREC. In *Text REtrieval Conference* (pp. 21–30.

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–9–5). Oxford University Press.

Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of SIGIR-1999* (pp. 279–280.