HOENKAMP, E., BRUZA, P., HUANG, Q. and SONG, D. 2008. The asymptotic behavior of a limited dependencies language model. In *Hoenkamp, E., De Cock, M. and Hoste, V. (eds.) Proceedings of 8th Dutch-Belgian information retrieval workshop 2008 (DIR 2008), 14-15 April 2008, Maastricht, Netherlands*. Enschede: Neslia Paniculata, pages 59-64.

The asymptotic behavior of a limited dependencies language model.

HOENKAMP, E., BRUZA, P., HUANG, Q. and SONG, D.

2008

The Asymptotic Behavior of a Limited Dependencies Language Model

Eduard Hoenkamp
University of Maastricht
The Netherlands
hoenkamp@acm.org

Peter Bruza
Queensland University of Technology
Australia
p.bruza@qut.edu.au

Qiang Huang
Open University
United Kingdom
q.huang@open.ac.uk

Dawei Song
Open University
United Kingdom
d.song@open.ac.u

ABSTRACT

Intuitively, any 'bag of words' approach in IR should benefit from taking term dependencies into account. Unfortunately, for years the results of exploiting such dependencies have been mixed or inconclusive. To improve the situation, this paper shows how the natural language properties of the target documents can be used to transform and enrich the term dependencies to more useful statistics. This is done in three steps. First, the term co-occurrence statistics of queries and documents are each represented by a Markov chain. The paper proves that such a chain is ergodic, and therefore its asymptotic behavior is unique, stationary, and independent of the initial state. Second, the stationary distribution is taken to model queries and documents, rather than their initial distributions. Third, ranking is achieved by comparing the Kullback-Leibler divergence between the stationary distributions of query and documents. These steps can be implemented as a simple and computationally inexpensive algorithm. The main contribution of this paper is to argue why the asymptotic behavior of the document model is a better representation of the document than any model that represents the dependencies in the document by its initial distribution. A secondary contribution is to investigate the practical application of this representation. To do so, the algorithm was tested on the AP88-89 and WSJ87-92 collections in a pseudo-relevance feedback setting. Results showed consistent improvements over a standard language model baseline. Moreover, even in its simple form, the algorithm proved already to be on a par with more sophisticated algorithms that depend on choosing sets of parameters or extensive training. Hence, adding such schemes may be expected to improve the the results of the simple algorithm beyond current practice.

1. INTRODUCTION

Imagine (or perhaps recall) that you just came back from a well-deserved vacation in the South Pacific. When someone asks you about your vacation, you are happy to recount how it was. First you tell it to the people at home, then to your neighbors, then to your colleagues at work. At first there will be much variation in your story, but by and by all has been said, and the rendition of your experience becomes stable, only mentioning the essential parts. Or think of an event that lands as late breaking news on your paper's front page. As days go by, the story may reappear a few times, but eventually all has been said.

Now suppose a search engine would need to return the most relevant (as opposed to the most entertaining) story about your vacation. Should it be one from the earlier stages where it still meandered haphazardly along all that happened? Or one of the later more concise and orderly accounts?

Let us look at this phenomenon from the language modeling perspective to IR [10]. In this paradigm a text is viewed as a sample from a stochastic source that produces words according to some distribution. With the vacation story, you were the source, and your stories were different samples from that source. As the source is assumed to be stochastic, the words and their frequencies will change from one account to the next, as in the case of your stories.

Without a model of the underlying process, however, it would be difficult to reconstruct the distribution of the source from the samples alone. Therefore, language models can be distinguished by how they model the source and by how the distribution is derived from the samples. As current language models don't use an explicit representation of the meaning of documents, we can illustrate our approach with a simple abstract example. Assume a language of just the words a and b, and two documents $D_1 = [a\ a\ a\ a\ a\ b\ b\ b\ b\ b\ b\ a]$ and $D_2 = [a\ b\ a\ b\ a\ b\ a\ b\ a\ b\ a\ b]$. Using $Q = [a\ b\ a\ b\ ]$ as the query (or topic), which document would be considered the most relevant for a given language model? In the multi-bernoulli model [10], D_1 and D_2 would get the same score, as all words in the query are also in the documents. The multinomial unigram model [15] also assigns the same score because the frequencies of a and b are the same in D_1 and D_2 and hence the $p(Q|D) = \prod_i p(q_i|D)$ are the same. If Q were extended with a word c that does not appear in the documents, so that smoothing [17] was called for, words would be discounted by the same amount, and again the documents would receive the same score. Basically, we are trying to estimate a relevance model (1) without further knowledge about the corpus, (2) under the assumption that the term occurrences are independent, and (3) in the absence of training data. These issues have received much attention lately. For example, several researchers have stud-

ied bigrams and trigrams [15] or even studied the optimal distance over which to consider dependencies in general [14, 9] or based on natural language constraints [5]. Metzler and Croft [9] in particular distinguished among full independence, sequential dependence, and full dependence. The terms mean what they suggest: in sequential dependence the ranking of a document depends only on the dependency of adjacent words, whereas in full dependence any clique of words is to be considered. In this paper we consider a fourth option, halfway between sequential and full dependence, namely when a word comes after another, but separated by words in between. For example, in D_1 and D_2 above, one can accumulate the distances from every a to every b to derive a probability that a is followed by b. In the example, this probability is much lower for D_1 than for D_2. Imagine that, as in the vacation story that was told over and over again, the sources of D_1 and D_2 would go on for a long time producing one new document after another according to their distributions. If we assume for concreteness a dependency of no more than five words, then (as we will see) in the long run a would appear about as often as b for D_2 but twice as often for D_1. This is obviously different from the word counts that would suggest a 50% probability for each. Moreover, the distribution in the long run seems to reflect the impression that D_2 is more like Q than is D_1. This paper will show how the term dependencies of a particular document predict the asymptotic behavior of its source, and with it the term distribution that would be observed if the source would continue to produce new documents.

The sections that follow show how the approach of asymptotic behavior relates to other language models, and how it accomplishes the following objectives:

- It shows that under very realistic, plausible, and elementary conditions the *source underlying a document is ergodic*, and therefore a stationary distribution to represent the source can be derived from just one document,
- It shows how documents can be ranked based on their underlying stationary distributions.
- It shows how an initial (ad hoc) distribution for a document can be establish , based on a semantic approach called the *Hyperspace Analog to Language* (HAL),

2. THE DOCUMENT SOURCE AS AN ERGODIC CHAIN

One reason that language models use lower order dependencies is the (in)tractability of the Bayesian chain rule. Another is often simply a lack of knowledge about higher order dependencies. Yet, in practice, bigrams already give a reasonable improvement over unigrams [6]. In addition, [15] and others have shown that an interpolation of unigram and bigram models performs well.

The practical considerations aside, the question remains whether higher order dependencies would lead to better models, even if it is tempting to assume the affirmative. To begin answering the question, it is important to realize that the current approach to language modeling is applicable to any stochastic source and the languages they produce (human, machine, or perhaps alien). The models pay no heed to the fact that the documents to be modeled are produced by humans. Yet this throws out particular constraints that could make the methods more tractable. Some constraints can be borrowed from cognitive science, some follow directly from confining the languages under consideration to natural language:

- Many cognitive phenomena can be understood sufficiently well in terms of word-pairs. Pertinent examples can be found e.g. in the research on memory [14], work as mentioned above on the 'semantic space' [2], and results from old theories on 'spreading activation' [1] to recent brain (ERP) studies [4]. This supports the view that the source underlying the document can be modeled as a (first order) Markov process.
- 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. Identifying words with the states of the process then means that 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*.

The first point was already proposed by Shannon in his famous article [13], without the backup from cognitive science. The next two points, that the Markov process is both aperiodic and irreducible means that it is *ergodic*. An ergodic chain has the property that in the long run it reaches a stationary distribution (also called stationary kernel, or steady state), irrespective of the initial state.
It is easy to sample a document and generate a new one on the basis of its distribution; see the examples in [13], or any of the many sites on the web that offer programs to do this. What we would like to compute however is the distribution of the source underlying the document. Or in the metaphor of the introduction, we would like to model the final stable and concise story as the most relevant to the query about the vacation. With little knowledge of the source, one could use a Gibbs sampler, i.e. generate a long series of documents and sample until the distribution seems to converge. The Gibbs sampler was proposed for example by Wei and Croft [16] to find a distribution for their LDA model. Besides the benefits of the technique, there are several issues to overcome: (1) it is computationally demanding, (2) it is hard to know when the process has converged, and (3) it is not certain if the outcome is the only fixed point. The observation above that the process is ergodic obviates all three issues at once. The final distribution of the Markov chain can easily be computed without sampling (it is the eigenvector with eigenvalue 1), and it is guaranteed to be unique.

Note, first, that the properties mentioned to derive this result are valid for natural languages in general. This means that the method may be used for languages other than English (and which are increasingly visible on the Web). Second, it also answers the question about the higher order dependencies, in that it is unlikely that these will contribute much to improving search results. With the answer comes an other question to the fore: how to compute the lower order dependencies given the documents. The next section offers a proposal, one we will use in an experiment further on, but it is by no means meant as the last word on finding initial distributions.

3. DERIVING THE INITIAL DISTRIBUTION

In language modeling, the document source represents the author producing the document. As an author could produce different renderings of the same story, these renderings would be different samples of the source, and so the term distribution could differ from one document to the next. Fortunately, the ergodic chain has a property that is very useful here, namely that its asymptotic behavior is independent of the initial state. In other words, if one would continue to sample the source, then in the long run it would not matter what sample, i.e. what document, was observed first; the asymptotic behavior would be the same. What remains then, is to derive an initial distribution given the document. This is where language models differ greatly from one another. As we mentioned in the introduction, an important distinction lies in the degree of term dependency that is assumed. In this paper we follow the approach of Lund and Burgess [8] who computed co-occurance statistics from a rich source of spontaneous conversations: Usenet newsgroups. They called the representation of these statistics the 'Hyperspace Analog to Language' or HAL. HAL is computed by sliding a window over the corpus and assigning weights to word pairs, inversely to the distance from each word to every other in the window. This results in a word by word matrix with the accumulated word distances in the cells. **Box 1** may clarify the construction further.

Box 1

Given an n-word vocabulary, the HAL space is represented as a $n * n$ matrix constructed by moving a window of size w over the corpus ignoring punctuation, sentence, and paragraph boundaries. The strength of co-occurence decreases with the number of intervening words. Instead of an extended corpus, let us take just the sentence *The effects of spreading pollution on the population of Atlantic salmon.*

	the	effects	of	spreading	pollution	on	population	atlantic	salmon
the		1	2	3	4	5			
effects	5								
of	8	5		1	2	3	5		
spreading	3	4	5						
pollution	2	3	4	5					
on	1	2	3	4	5				
population	5		1	2	3	4			
atlantic	3		5		1	2	4		
salmon	2		4			1	3	5	

The table above shows the HAL matrix for a window size of 5. Take e.g. the entry for 'population'. To find the distance to 'pollution', go backward starting at 'population' with strength 5 (for 'the') counting down to 3 for 'pollution'.

(Note that the number in a cell is formally not a distance because the matrix is usually not symmetric.) If a word is connected to a second word via a small number, than it is more likely to be followed by that word than if the number had been high (e.g. the table shows that 'of' is more likely to be followed by 'the' than the other way around). Based on this observation, the HAL matrix is transformed into a transition probability matrix *pHAL* by normalizing the row vectors. So, to find the document source distribution for a document requires only two steps:

1. Compute the ad-hoc distribution, in our case pHAL,
2. Compute the stable distribution (epi-HAL).

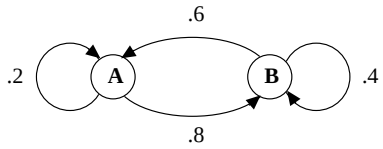
epi-HAL, for 'ergodic process interpretation of HAL', is easy to compute in several ways, which follow from the ergodic property. For example, one can compute the eigenvector of pHAL that belongs to the eigenvalue of 1.

Box 2

For readers unfamiliar with the Markov approach, the essential steps in the algorithm are illustrated below. Assume a language of just the words a and b, whose dependencies are defined by the transition probabilities in matrix H. H defines a Markov chain, where state **A** ouputs a and state **B** outputs b.

H	a	b
a	.2	.6
b	.8	.4

$\Rightarrow$



For initial state s_0 (e.g. **A** if started with word a), 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

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

which converges to:

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

so the Markov chain becomes stationary with $P(a) = .4286$ and $P(b) = .5714$, independent of the initial state. (The formal derivation was only given to show the convergence. The stationary distribution can also be computed directly from the transition matrix.) In the same way these values can be obtained for the examples in the introduction. Computing the HAL matrix with window of size 4, the distributions converge to:
$D_1 = [a\ a\ a\ a\ a\ b\ b\ b\ b\ b\ b\ a]$, $P(a) = .36$ and $P(b) = .64$
$D_2 = [a\ b\ a\ b\ a\ b\ a\ b\ a\ b\ a\ b]$, $P(a) = .49$ and $P(b) = .51$
$Q = [a\ b\ a\ b\ ]$, $P(a) = .44$ and $P(b) = .56$
Computing the Kullback-Leibler divergence yields
$KL(Q||D_1) = .017$, and $KL(Q||D_2) = .007$, so D_1 diverges more from Q than D_2, and therefore D_2 is ranked as more relevant.

Doing this for all documents produces a source representation for each document. The same can be done for the

query, which would represent the searcher. To rank the documents in order of relevance to the searcher, the documents are not compared to the query directly (as in the vector space model) but the sources are compared. Researchers in the language modeling community use the Kullback-Leibler (KL) divergence to compare distributions, and so will we. The algorithm is explained in **Box 2** using a very simple language for clarity.

The main goal of this paper is to explain and more formally justify our approach. Yet, the next section will add a more practical justification by showing that even a straightforward and simple implementation of our approach can already compete with a closely related but much more sophisticated language model.

4. IMPLEMENTATION AN EVALUATION

There are other language models that use a Markov approach. Notably Cao, Nie, and Bai [3] use the Markov chain for the same reason as we do, namely to find a stable distribution. There are a number of choices made in [3] that we do not depend on: we do not use WordNet (for semantic relationships), there are several parameters we do not have to set, and we don't use training for optimization. Furthermore, although in the authors of [3] make use of a stationary distribution, they cannot guarantee that their initial distribution is ergodic. So their stationary points may not be unique and depend on the initial distribution, i.e. on the precise rendering of the document. At the time of writing we have not yet compared our results with theirs, so we will report our findings once we have done the experiment. We did, however, conduct an experiment to compare our model with the somewhat older relevance model of Lavrenko & Croft [7].

The evaluation experiment follows a pseudo-relevance feedback paradigm, which requires a few choices to be made which we will mention here. This makes the model less elegant, but it was necessary in order to evaluate the approach against the work of others.

First a document ranking is produced in response to a query Q. The top n documents are used to derive a distribution M^n_{epi} by computing the epi-HAL over this collection. Similarly, M^Q_{epi} is computed for the query. These are used in turn to define a mixture model (cf. equation (15) in [7]).

$$\Pr(w|Q) = \lambda \Pr(w|M^Q_{\text{epi}}) + (1-\lambda)\Pr(w|M^n_{\text{epi}}) \qquad (1)$$

The documents were re-ranked using the KL-divergence, and we used the standard baseline unigram LM in the Lemur toolkit. In the experiments reported below a simplified version Robertson's term selection value (TSV) worked well in the case of HAL, and which was defined as

$$\text{WT}(w) = \frac{r_w}{R} \log \frac{N}{f_w} \qquad (2)$$

where f_w is the occurrence frequency of word w within a corpus of N documents, R is the number of selected top-ranked documents and r_w is the number of documents that contain a particular term w. We established the number of terms by balancing the number of relevant terms and providing sufficient dimensionality for pHAL. We found a value of 300 terms, but others have used different values here. (Such differences are to be expected as the distributions are calculated differently, and there is no better way known then to establish these numbers empirically.) Once this number is established it can be used to compute $\Pr(\cdot|M^n_{\text{epi}})$ and $\Pr(\cdot|M^Q_{\text{epi}})$. Substituting these in equation (1) yields the query model $M_Q = \Pr(w|Q)$. Subsequently, documents are re-ranked via $KL(M_Q||M_D)$, where M_D corresponds to a document language model. In our case, M_D is delivered by the baseline language model.

4.1 Experimental Results

We will now compare the present proposal to the relevance model. We used the TREC corpora Associated Press 88-89 and Wall Street Journal 90-92 used in [7]. The results are tabulated below together with a measure of robustness. It can be seen from the robustness data that the epi-HAL approach improves substantially over the baseline, but that the difference with RM is negligible. The same happens in Table 2. For the Wall Street Journal, we see a significant increase in performance of the epi-HALL approach over RM, for all topics.

Table 1: Comparison of mean average precision (MAP) by testing Query101-150 on Collection AP8889, using KL as baseline, the relevance model RM, and epi-HAL proposed in the present paper. The robustness numbers below the table give an impression of how consistent the differences are, if any.

	KL	RM	epi-HAL	ΔKL %	ΔRM %
MAP	0.2336	0.3047	0.3089	+32.2**	+1.4
Recall	3179	3919	3952		

Robustness EMC vs. KL: Positive 31, Negative 18
Robustness EMC vs. RM: Positive 23, Negative 26
Robustness RM vs. KL: Positive 34, Negative 15

Table 2: See Table 1, AP8889, Query151-200

	KL	RM	epi-HAL	ΔKL %	ΔRM %
MAP	0.3084	0.3794	0.3807	+23.4**	+0.3
Recall	3332	3636	3697		

Robustness EMC vs. KL: Positive 33, Negative 16
Robustness EMC vs. RM: Positive 21, Negative 28
Robustness RM vs. KL: Positive 32, Negative 17

Table 3: See Table 1, WSJ90-92, Query101-150

	KL	RM	epi-HAL	ΔKL %	ΔRM %
MAP	0.2568	0.2632	0.2836	+10.4**	+7.7*
Recall	1537	1464	1606		

Robustness EMC vs. KL: Positive 30, Negative 18
Robustness EMC vs. RM: Positive 31, Negative 17
Robustness RM vs. KL: Positive 25, Negative 23

Table 4: See Table 1, WSJ90-92, Query151-200

	KL	RM	epi-HAL	ΔKL %	ΔRM %
MAP	0.2336	0.3047	0.3089	+32.2	+1.4
Recall	3179	3919	3952		

Robustness EMC vs. KL: Positive 31, Negative 18
Robustness EMC vs. RM: Positive 23, Negative 26
Robustness RM vs. KL: Positive 34, Negative 15

5. CONCLUSION

We derived a relatively simple language model, epi-HAL, that deviates in several respects from most language models proposed to date. Epi-HAL is based on the observation that texts are produced by humans. From this observation it follows that (1) there must be semantic dependencies underlying the documents, and (2) that the documents must obey surface constraints inherent to natural language. To represent the former, this paper derived the underlying semantics from the Hyperspace Analog to Language (HAL) a theory presuming that words that appear close together in text, will also be close in meaning. The surface constraints were represented by using an ergodic Markov chain.

We believe that current language models are overly general in that they do not incorporate these properties of natural language, the very fabric of the documents they purport to model. We compared the straightforward implementation of the proposed model with a sophisticated relevance model. Evaluation on TREC corpora showed results that are on a par with with the relevance model in the case of AP8889, and in the case of WSJ90-92 even performed better.

The results of the experiments encourages us to pursue several avenues in future work. First, the current experiments can be repeated for comparison with other Markov approaches, such as the recent [3] which makes a number of additional assumptions that we don't have to make. Second, we could apply the ergodic chain approach to replace the maximum likelihood document model in [7] by the stationary distribution. And finally, because the proposed model itself is so simple there is always room to improve results via specialized additions as are currently used in much more sophisticated models.

6. REFERENCES

[1] J. Anderson. *The Architecture of Cognition*. Harvard University Press, Cambridge, MA, USA, 1983.

[2] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25:211 – 257, 1998.

[3] G. Cao, J. Y. Nie, and J. Bai. Using markov chains to exploit word relationships in information retrieval. In *the 8th Conference on Large-Scale Semantic Access to Content (RIAO07)*, 2007.

[4] D. Chwilla and H. Kolk. Accessing world knowledge: Evidence from n400 and reaction time priming. *Cognitive Brain Research*, 25:589–606, 2005.

[5] J. Gao, J.-Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 170–177. ACM Press, 2004.

[6] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for IR. In *Proceedings of the 24th Conference on Research and Development in Information Retrieval*, pages 111–119, 2001.

[7] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, 2001.

[8] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.

[9] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2005. ACM Press.

[10] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

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

[12] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30, 1992.

[13] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[14] R. M. Shiffrin and M. Steyvers. The effectiveness of retrieval from memory. In M. Oaksford and N. Chater, editors, *Rational models of cognition*, pages 73–9–5. Oxford University Press, 1998.

[15] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of the 22nd Conference on Research and Development in Information Retrieval*, pages 279–280, 1999.

[16] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, USA, 2006. ACM Press.

[17] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, New York, NY, USA, 2001. ACM Press.