Learning and Representing Topic
A Hierarchical Mixture Model for Word Occurrences in Document Databases

Thomas Hofmann
Center for Biological and Computational Learning, MIT
Cambridge, MA 02139, USA, hofmann@ai.mit.edu

Abstract
This paper presents a novel statistical mixture model for natural language learning in information retrieval. The described learning architecture is based on word occurrence statistics and extracts hierarchical relations between groups of documents as well as an abstractive organization of keywords. To train the model we derive a generalized, annealed version of the Expectation–Maximization (EM) algorithm for maximum likelihood estimation. The benefits of the model for interactive information retrieval and automated cluster summarization are experimentally investigated.

1 Introduction
Intelligent processing of text and documents ultimately has to be considered as a problem of natural language understanding. In this paper, I will present a statistical approach to learning of language models for context–dependent word occurrences and discuss the applicability of this model for interactive information retrieval. The proposed technique is purely data–driven and does not make use of domain–dependent background information, nor does it rely on predefined document categories or a given list of topics. The presented cluster–abstraction model (CAM) is a statistical mixture model [6, 5] which organizes groups of documents in a hierarchy. Compared to most state–of–the–art techniques based on agglomerative clustering (e.g., [4, 1, 9]) is has several advantages and additional features As a generative model the most important advantages are: (i) a sound foundation on the likelihood principle (likelihood as a global clustering criterion), (ii) the probabilistic inference mechanism, (iii) evaluation of generalization performance for model complexity control, (iv) efficient model fitting by the EM algorithm, (v) explicit representation of conditional independence relations. Additional advantages are provided by the hierarchical nature of the model, namely: (vi) multiple resolution levels of document clustering, (vii) discriminative topic descriptors for document groups, (viii) coarse–to–fine approach by deterministic annealing.

2 Probabilistic Clustering of Documents
Let us emphasis the clustering aspect by first introducing a simplified, non–hierarchical version of the CAM which performs ‘flat’ probabilistic clustering and is closely related to the model proposed in [7] for word clustering. Let the symbols $d_i$ ($1 \leq i \leq N$) and $w_j$ ($1 \leq j \leq M$) denote documents and words (word stems), respectively. Counts for word $w_j$ in document $d_i$ are denoted by $n_{ij}$ and $n_i = \sum_j n_{ij}$ is the total number of words in document $d_i$. Following the standard mixture approach, it is assumed that each document belongs to one out of $K$ clusters $C_\alpha$. These hidden variables are represented by indicator functions $H_{i\alpha} \in \{0,1\}$, i.e., $H_{i\alpha} = 1$ if $d_i$ belongs to $C_\alpha$. Moreover, let us introduce
parameters $p_{j|a} = P(w_j|C_a)$ for cluster-specific word probability distributions. Then we can specify a joint probability model by:

$$P(d_i, H_{ia} = 1|p, \pi) = \pi_a \prod_{j=1}^{M} (p_{j|a})^{n_{ij}},$$

where $\pi_a$ are parameters for the prior distribution of $H_{ia}$. The factorial expression for the joint probability reflects conditional independence assumptions about word occurrences (bag-of-words model). Starting from (1) the standard EM approach [2] yields the following coupled re-estimation equations

$$P(H_{ia} = 1|p, \pi) = \frac{\pi_a \prod_{j=1}^{N} (p_{j|a})^{n_{ij}}}{\sum_{a=1}^{K} \pi_a \prod_{j=1}^{N} (p_{j|a})^{n_{ij}}},$$

$$\pi_{a}^{new} = \frac{1}{N} \sum_{i=1}^{N} P(H_{ia} = 1|p^{old}, \pi^{old}), \quad p_{j|a}^{new} = \frac{\sum_{i=1}^{N} P(H_{ia} = 1|p^{old}, \pi^{old}) n_{ij}}{\sum_{i=1}^{N} P(H_{ia} = 1|p^{old}, \pi^{old}) n_{i}}.$$

These equations are very intuitive: The posteriors encode a probabilistic clustering of documents, while the conditionals $p_{j|a}$ represent average word distribution for documents belonging to group $C_a$. Of course, the simplified flat clustering model defined by (1) has several deficits. Most severe are the lack of a multi-resolution structure and the inadequacy of the ‘prototypical’ distributions $p_{j|a}$ to emphasis discriminative or characteristic words (they are in fact dominated by the most frequent word occurrences). To cure this flaws is the task of the hierarchical extension presented in the next section.

## 3 Document Hierarchies and Abstraction

Most hierarchical document clustering techniques utilize agglomerative algorithms which generate a cluster hierarchy or dendogram as a by-product of successive cluster merging. In the CAM we will use an explicit abstraction model instead to represent hierarchical relations between document groups. This is achieved by extending the ‘horizontal’ mixture model of the previous section with a ‘vertical’ component that captures the specificity of a particular word $w_j$ in the context of a document $d_i$. It is thus assumed that each word occurrence $(d_i, w_j)$ was generated from an abstraction level $\mathcal{A}_v$, where abstraction levels are identified with inner or terminal nodes of the cluster hierarchy (cf. Figure 1 (a)).

To formalize the sketched ideas, additional hidden variables $V_{i,j,v} \in \{0, 1\}$ are introduced for each word occurrence with $V_{i,j,v} = 1$ if $(d_i, w_j)$ was generated from $\mathcal{A}_v$. The hidden variables have to fulfill the following sets of constraints: $\sum_{a} \sum_{\mathcal{A}_v \uparrow C_a} H_{ia} V_{i,j,v} = 1$, where $\mathcal{A}_v \uparrow C_a$ denotes the nodes $\mathcal{A}_v$ ‘above’ $C_a$, i.e., nodes on the path to $C_a$. A pictorial representation can be found in Figure 1 (b): if $d_i$ is assigned to $C_a$ the choices for abstraction levels of occurrences are restricted to the ‘active’ (highlighted) vertical path.

Generalizing the non–hierarchical model, a probability distribution $p_{v|w}$ over words is attached to each node (inner or terminal) of the hierarchy. The complete data model is given by

$$P((d_i, w_j), H_{ia} = 1, V_{i,j,v} = 1|p, \pi, \rho) = \pi_a p_{v|w, i, a} p_{j|v},$$

where the additional $\rho$ parameters are prior probabilities for $V$ ($p_{v|w, i, a} = 0$ whenever $\mathcal{A}_v \not\uparrow C_a$ in the given tree). The prior probabilities $p_{v|w, i, a}$ capture document-specific distribution over abstraction levels (conditioned on the fact that $d_i$ belongs to $C_a$). Marginalization over the hidden variables results

---

3For simplicity the number of words in a document is not treated as a random variable and assumed to be given a priori.
in the following log-likelihood of the ‘double’ mixture model

\[ \mathcal{L} = \sum_{i=1}^{N} \log \sum_{\alpha=1}^{K} \pi_{\alpha} \prod_{j=1}^{M} \left( \sum_{\nu} \rho_{\nu|\alpha}(i,\nu) P_{j|\nu} \right)^{n_{ij}} . \] (5)

As for the simplified model before, we will derive an EM algorithm for model fitting. The E–step requires to compute (joint) posterior probabilities of the form \( P(H_{i\alpha}|V_{(i,j)\nu} = 1|p^{old}, \pi^{old}, \rho^{old}) \) (abbreviated by \( q_{i\alpha}^{ij} \) in the sequel). After decomposing by the chain rule one obtains

\[ P(V_{(i,j)\nu} = 1|H_{i\alpha} = 1, p, \rho) = \frac{\rho_{\nu|\alpha}(i,\nu) P_{j|\nu}}{\sum_{\mu} \rho_{\mu|\alpha}(i,\mu) P_{j|\mu}} \quad P(H_{i\alpha} = 1|p, \pi) \propto \pi_{\alpha} \prod_{j=1}^{M} \left( \sum_{\nu} \rho_{\nu|\alpha}(i,\nu) P_{j|\nu} \right)^{n_{ij}} . \] (6)

The M–step re-estimation equations are given by

\[ P_{j|\nu}^{new} = \frac{\sum_{i=1}^{N} \sum_{\alpha=1}^{K} q_{\alpha}^{ij} n_{ij}}{\sum_{i=1}^{N} \sum_{\alpha=1}^{K} q_{\alpha}^{ij} n_{i}}, \quad P_{\nu|\alpha}^{new}(i,\alpha) \propto \sum_{j=1}^{M} q_{\alpha}^{ij} n_{ij}, \quad \pi_{\alpha}^{new} \propto \sum_{i=1}^{N} P(H_{i\alpha} = 1|p_{\alpha}^{old}, \pi_{\alpha}^{old}) . \] (7)

Finally, it may be worth taking a closer look at the predictive word probability distribution \( p_{j|i} \) in the CAM which is given by \( p_{j|i} = \sum_{\alpha} P(H_{i\alpha} = 1|p, \pi) \sum_{\nu} \rho_{\nu|\alpha}(i,\nu) P_{j|\nu} \). If we assume for simplicity that \( P(H_{i\alpha} = 1|p, \pi) = 1 \) for some \( \alpha = \alpha_0 \), then the word probability of \( d_i \) is modeled as a mixture of occurrences from different abstraction levels \( A_\alpha \). This reflects the reasonable assumption that each document contains a certain mixture of words ranging from general terms of ordinary language to highly specific technical terms and specialty words.

There are three important problems which need also to be addressed in a successful application of the CAM: First, one has to avoid the problem of overfitting. Second, it is necessary to specify a method to determine a meaningful tree topology including the maximum number of terminal nodes. And third, one may also want to find ways to reduce the sensitivity of the EM procedure to local maxima. An answer to all three questions is provided by a generalization called annealed EM [3]. Annealed EM is closely related to a technique known as deterministic annealing that has been applied to many clustering problems (e.g. [8, 7]). Since a thorough discussion of annealed EM is beyond the scope of this paper, I will skip the theoretical background and focus on a procedural description. The key idea in deterministic annealing is the introduction of a temperature parameter \( T \in [0, 1] \). Applying the annealing principle to the horizontal hidden variables \( H \) the corresponding posterior calculation in (6) is generalized by replacing \( n_{ij} \) in the exponent by \( n_{ij}/T \). Downweighting the number of observations
Figure 2: (a) Abstract from the generated LEARN document collection, (b) representation in terms of word stems, (c) words with lowest perplexity under the CAM for words not occurring in the abstract (differentiated according to the hierarchy level).

<table>
<thead>
<tr>
<th>Most frequent words</th>
<th>CAM most top words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>#33</strong></td>
<td><strong>#25</strong></td>
</tr>
<tr>
<td>learn</td>
<td>learn</td>
</tr>
<tr>
<td>network</td>
<td>example</td>
</tr>
<tr>
<td>neural</td>
<td>algorithm</td>
</tr>
<tr>
<td>algorithm</td>
<td>weight</td>
</tr>
</tbody>
</table>

Figure 3: Group descriptions for exemplary inner nodes by most frequent words and by the highest probability words from the respective CAM node.

by taking the likelihood contribution to the $1/T$-th power for $T > 1$ emphasizes the prior and will in general increase the entropy of the (annealed) posterior probabilities. In annealed EM, $T$ is utilized as a control parameter which is initialized at a high value and successively lowered until the performance on a held-out set starts to decrease. Annealing is advantageous for model fitting, since it offers a simple and inexpensive way to control the effective model complexity. This avoids overfitting and improves the average solution quality of EM procedures. Moreover, it also offers a way to generate tree topologies, since annealing leads through a sequence of so-called phase transitions. More details on this subject can be found in [3].

4 Results

All documents utilized in the experiments have been preprocessed by word suffix stripping with a word stemmer. A standard stop word list has been utilized to eliminate the most frequent words, in addition very rarely occurring words have also been eliminated. An example abstract and its index term representation is depicted in Figure 2 (a), (b). The experiments reported are some spotlights selected from a much larger number of performance evaluations. They are based on two datasets which form the core of our current prototype system: a collection of 3609 recent papers with ‘learning’ as a titleword, including all abstracts of papers from Machine Learning Vol. 10-28 (LEARN), and a dataset of 1568 recent papers with ‘cluster’ in the title (CLUSTER).

The first problem we consider is to estimate the probability for a word occurrence in a text based on the statistical model. Figure 2 (c) shows the most probable words from different abstraction levels, which did not occur in the original text of Figure 2 (a). The abstractive organization is very helpful to distinguish layers from trivial suggestions of unspecified word occurrences up to highly specific technical
One of the most important benefits of the CAM is the resolution-specific extraction of characteristic keywords. In Figure 4 we have visualized the top 6 levels for the dataset LEARN. The overall hierarchical organization of the documents is very satisfying, the topological relations between clusters seems to capture important aspects of the inter-document similarities. In contrast to most multi-resolution approaches the distributions at inner nodes of the hierarchy are not obtained by a coarsening procedure which typically performs some sort of averaging over the respective subtree of the hierarchy. The abstraction mechanism in fact leads to a specialization of the inner nodes. This specialization effect makes the probabilities \( p_{\text{LP}} \) suitable for cluster summarization. Notice, how the low-level nodes capture the specific vocabulary of the documents associated with clusters in the subtree below. The specific terms become automatically the most probable words in the component distribution, because higher level nodes account for more general terms. To stress this point we have compared the abstraction result with probability distributions obtained by averaging over the respective subtree. Figure 3 summarizes some exemplary comparisons showing that averaging mostly results in high probabilities for rather unspecific terms, while the CAM node descriptions are highly discriminative. The node-specific word distribution thus offer a principled and very satisfying solution to the problem of finding resolution-specific index terms for document groups as opposed to many circulating ad hoc heuristics to distinguish between typical and topical terms.
Figure 5: Example run of an interactive image retrieval for documents on ‘texture-based image segmentation’ with one level look-ahead in the CAM hierarchy.

An example run for an interactive coarse-to-fine retrieval with the CLUSTER collection is depicted in Figure 5, where we pretend to be interested in documents on clustering for texture-based image segmentation. In a real interactive scenario, one would of course display more than just the top 5 words to describe document groups and use a more advanced shifting window approach to represent the actual focus in a large hierarchy. In addition to the description of document groups by inner node word distributions, the CAM also offers the possibility to attach prototypical documents to each of the nodes (the ones with maximal probability \( p_{e|t} \)), to compute most probable documents for a given query, etc. All informations, the cluster summaries by (locally) discriminant keywords, the keyword distributions over nodes, and the automatic selection of prototypical documents are particularly beneficial to support an interactive retrieval process. Due to the abstraction mechanism the cluster summaries are expected to be more comprehensible than descriptions derived by simple averaging. The hierarchy offers a direct way to refine queries and can even be utilized to actively ask the user for additional specifications.

Conclusion: The cluster- abstraction model is a novel statistical approach to natural language learning for information retrieval which has a sound foundation on the likelihood principle. The dual organization of document cluster hierarchies and keyword abstractions makes it a particularly interesting model for interactive retrieval. The experiments carried out on small/medium scale document collections have emphasized some of the most important advantages. Since the model extracts hierarchical structures and supports resolution dependent cluster summarizations, the application to large scale databases seems promising.

References


