massachusetts institute of technology — artificial intelligence laboratory

Model Selection in Summary Evaluation

Luis Perez-Breva and Osamu Yoshimi

AI Memo 2002-023 December 2002
CBCL Memo 222

Abstract

A difficulty in the design of automated text summarization algorithms is in the objective evaluation. Viewing summarization as a tradeoff between length and information content, we introduce a technique based on a hierarchy of classifiers to rank, through model selection, different summarization methods. This summary evaluation technique allows for broader comparison of summarization methods than the traditional techniques of summary evaluation. We present an empirical study of two simple, albeit widely used, summarization methods that shows the different usages of this automated task-based evaluation system and confirms the results obtained with human-based evaluation methods over smaller corpora.[1]

[1] This report describes research done at the Center for Biological & Computational Learning, which is in the Department of Brain & Cognitive Sciences at MIT and which is affiliated with the McGovern Institute of Brain Research and with the Artificial Intelligence Laboratory.

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, and National Science Foundation-NIH (CRCNS) Contract No. IIS-0085836.

Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., The Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sumitomo Metal Industries, Toyota Motor Corporation, WatchVision Co., Ltd., and The Whitaker Foundation.

1 Introduction

Evaluation of text summaries can be divided into two main trends, those that attempt to evaluate the quality of the summaries ("ideal" based), and those that assess performance in a given task ("task-based"). As currently applied, both methods require human evaluation of the documents and the summaries. The ideal based evaluation regards human-made summaries as the target in the comparison with automatic summaries. However, as "there is no single correct summary" [4], different experts might summarize the same article differently therefore, "agreement between human judges becomes an issue" [8]. On the other hand, task-based methods require human evaluators to process the information in order to accomplish a task (often "categorization" or "reading comprehension"). In this case the difficulty lies in translating the results into ranking of the summaries. Both methods are thoroughly reviewed in the TIPSTER[2] project. Attempts to compare different methods using these evaluation tools concluded that improvements are required in order to establish differences between summarizers [10]. For the "ideal" based analysis Hongyan et al [10] suggested that the binary precision and recall measures are too sensitive to provide a robust evaluation, and, for the task-based analysis, the measures used do not translate to any indicative measure of quality.

Among the various attempts to improve these evaluation techniques, D. Marcu [14] emphasizes the need for evaluating not only the actual output (summaries) but the underlying assumptions of the method as well. In this paper, extending Marcu's position, we replace the *a posteriori* human judgment by an *a priori* judgment, and evaluate summaries by the amount and level of detail of the information from the original document they contain. Instead of evaluating summaries, we ask human experts to do two things, first define a set of labels that applies to the whole corpus of documents, and, second, describe the documents with this set of labels. We describe a new technique to evaluate summaries based on this labelling. We train a hierarchy of classifiers to become an expert in the evaluation of these labels for different representations of the documents, each a different summary, and use model selection to find the summarizer that describes best the original labelling conditional on user's preferences. The final result is a quantitative comparison between different summarizers in terms of information content and level of detail. Our results in comparing two simple summarization methods show that this procedure agrees with the results of previous work using less quantitative evaluation methods.

The paper is organized as follows: section 2 introduces the formalism and implementation of the evaluation system and its implementation. Section 3 introduces the methodology for performing controlled experiments on document summarizers and presents teh result of a comparison between lead-based and headline based sentence selection. Section 4 discusses the validity of the results obtained and the main critiques to this approach to summary evaluation.

2 An automated task-based evaluation system

Formalization

We describe each summary by a triple:

$$\langle length,\ information\ content,\ level\ of\ detail \rangle.$$

In order to evaluate information content and the level of detail, we define a strict partial order Q with binary relations $O = \{\sqsubset, \sqsupset, =\}$ over the level of detail of the questions $q_i \in Q$. Where $q_i \sqsubset q_j$ implies that $q_i \in Q$ is a question corresponding to the same topic as $q_j \in Q$ but with a higher level of detail. The partial order Q corresponds to a set of questions agreed upon by a set of human experts, and depends on the data and the coverage of the information desired. For each document in the corpus, a group of experts assigns a *yes/no* value to each question $q \in Q$. Accordingly, each document is characterized by a set of pairs

$$\left\{\langle q^{yes}_{(0)}, q^{no}_{(0)}\rangle, \langle q^{yes}_{(0,0)}, q^{no}_{(0,0)}\rangle, \langle q^{yes}_{(0,0,0)}, q^{no}_{(0,0,0)}\rangle, \langle q^{yes}_{(0,1)}, q^{no}_{(0,1)}\rangle, \cdots, \langle q^{yes}_{(1)}, q^{no}_{(1)}\rangle \cdots\right\}, \tag{1}$$

[2] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/

where $q_{(0)} \sqsupset q_{(0,1)}$, and $q_{(0)}$ and $q_{(1)}$ are not comparable, and $q^{yes}_{(0)}$ is the proportion of experts having answered *yes* to the question $q_{(0)}$. We call the characterization of a document $\mathbf{x_j}$ through equation (1) the *label* of the document, and identify it by the letter y_j. We further ask each human expert to ensure consistency in his/her judgment in such a way that the following condition holds:

$$q_i \ = \ no \implies q_\ell \ = \ no \quad \forall q_\ell \sqsupset q_i \tag{2}$$

Different groups of experts are called to evaluate the same set of questions Q on the same documents once summarized. The level of agreement between their judgment and the judgment of the first group of experts is an indication of the accuracy of the summary. The final user shall decide among the summarizers the one that provides the desired balance between the three parameters in the triple $\langle length,\ information\ content,\ level\ of\ detail\rangle$[3].

This viewpoint describes a hierarchical (task-based) categorization system (equation (1) can easily be translated into a tree-like graph through a Hasse diagram with a $\top$ element), in which documents are allowed to belong to several categories (branches in the hierarchy) and describe information through different levels of detail (depth in the hierarchy). This is a typical information retrieval situation.

Model Selection

The description of the task-based evaluation by a partial order Q and several groups of experts allows for an efficient implementation of an automated evaluation system using machine learning techniques. Once the corpus has been labeled by a group of human experts we partition the corpus into three sets: training ($\mathcal{T}$), 1^{st} validation ($\mathcal{V}_1$), and 2^{nd} validation ($\mathcal{V}_2$) (for the moment we make no assumptions about how the documents are represented in the corpus). A set of classifiers, each replying to a binary question q_i, is trained on $\mathcal{T}$ and its parameters are iteratively tuned on $\mathcal{V}_1$, to produce a hierarchy of classifiers. At this stage, this hierarchy can be considered as an expert in the evaluation of the label of documents belonging to this corpus.

Different summarizers produce different representations of the same corpus. Each can be used to train a hierarchy of classifiers. Measuring the performance of each hierarchy on the corresponding representation of the set $\mathcal{V}_2$ will translate into a ranking of the summarizers, closely related to the choice of $\langle length,\ information\ content,\ level\ of\ detail\rangle$. This procedure, namely, model selection through a validation set, is the same that we applied to tune the parameters of the classifiers with $\mathcal{V}_1$.

Implementation

We build a set of classifiers following the structure of the partial order Q. Classifiers $C^0_i, i = 1,..,m$ are trained to reply to each question $q_{(i)} \in Q, i = 1,..,m$ in the first level of the hierarchy. The classifiers are tuned with the 1^{st} validation set $\mathcal{V}_1$ to minimize the number of false negatives. Informally, this is equivalent to recast the classification problem as:

$$\big\{0 : Inconclusive, 1 : q_i = \text{``}clearly''no\big\}. \tag{3}$$

The word *clearly* implies that the classifier output will be 1 when its reply to the question q_i is *no* with a probability higher than $1-\delta$. In this case, the consistency requirement imposed on the human evaluators (2) makes processing $q_j \sqsubset q_i$ unnecessary. On the other hand, if the reply is *inconclusive*, there are two options:

- refine the result from C^0_i for question q_i with an additional classifier C^1_i,
- assume $q_i = yes$ and explore $q_j \sqsubset q_i$.

Formally, a classifier C^k_i that sees the document $\mathbf{x}_j$ with real label y_j applies the following decision criteria

$$f_{C^k_i}(\mathbf{x}) = \begin{cases} 1 & P(q_i = no \mid \mathbf{x}_j,\ \theta_{C^k_i}) > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

[3] Note that in this formalization, the concept of ideal summary has become a relative measure of the choice of the values of the triple $\langle length,\ information\ content,\ level\ of\ detail\rangle$. Thus allowing for different choices of summarizers based on user preferences.

where $\theta_{C_i^k}$ represents the parameters of classifier C_i^k, and the *Threshold* $\tau \in \mathbb{R}$ is set to ensure

$$P(y_j\{q_i\} = yes | f_{C_i^k}(\mathbf{x}_j) = 1) \leq \delta \;\; \forall \mathbf{x}_j \in \mathcal{V}_1. \tag{5}$$

The procedure for training classifiers for questions $q_j \sqsubset q_i$ is analogous to the one detailed for the first layer of the hierarchy.

Remarks

The labeling of the classes adopted in equation (3) is a direct consequence of the choice of δ in equation (5).The choice of a small value for δ makes classes in questions $q_j \sqsubset q_i$ more balanced and prevents errors from propagating across the structure. Similar strategies have been applied to real-time object detection in [20].

A side effect of the choice of a small δ is the high value of the quantity

$$P(y_j\{q_i\} = no | f_c(\mathbf{x_j}) = inconclusive). \tag{6}$$

In fact, the hierarchy is mostly concerned on discarding clearly negative samples. This may result in a high over-labeling[4] of each sample. In order to prevent over-labeling, additional classifiers C_i^j may be nested for each question q_i to evaluate the samples considered *inconclusive* by the classifier C_i^{j-1}.

Nesting classifiers has the effect of producing a two-dimensional hierarchy. The first dimension takes care of the depth of information, or, level of detail through the partial order Q; the second dimension is produced by nesting classifiers for each question q_i. It is responsible for the accuracy of the classifiers in their joint reply to the question q_i. In the limit, the two-dimensional hierarchy should yield a system that classifies samples with an accuracy bounded by δ, the depth of the hierarchy, and the number of nested classifiers per layer.

Evaluation measure

We benchmark performance on the hierarchy of classifiers using "soft" precision and recall measures [5] on the class probability output of each classifier. As noted before, and in [10], binary ("hard") precision and recall measures defined on the class output do not provide sufficient resolution.

Given the probability output $(P_{C_i^k})$ of the classifier C_i^k for question q_i on the sample $\mathbf{x}_j$, we define the following quantities:

$$M_j^i = \begin{cases} P_{C_i^k} & y_j\{q_i\} = yes, \\ 0 & \text{otherwise} \end{cases}, \quad U_j^i = \begin{cases} 1 - P_{C_i^k} & y_j\{q_i\} = yes, \\ 0 & \text{otherwise} \end{cases}, \quad O_j^i = \begin{cases} P_{C_i^k} & y_j\{q_i\} = no, \\ 0 & \text{otherwise} \end{cases}$$

Soft precision and recall of the labelling of sample x_j result from the equalities:

$$precision_{method}(x_j) = \frac{\sum_{q_i \in Q} M_j^i}{\sum_{q_i \in Q}(M_j^i + O_j^i)} \tag{7}$$

$$recall_{method}(x_j) = \frac{\sum_{q_i \in Q} M_j^i}{\sum_{q_i \in Q}(M_j^i + U_j^i)} \tag{8}$$

To compare visually different distributions we use quantile plots of the F-measure introduced in [19] and used in [8]:

$$F = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 precision + recall} \tag{9}$$

As noted in [1] for the usual choice $\beta = 1$, the F measure becomes the harmonic mean of the precision and recall distributions.

[4] We define over-labeling as assigning a *yes* value to questions q_i that are really negative in the original label y.

[5] the "soft" precision and recall measures defined here attempt to solve the same problems outlined in [8], but differ in the implementation.

User preferences

We now examine the role in the evaluation measure of the triple $\langle length,\ information\ content,\ level\ of\ detail\rangle$. User preferences can be introduced in the previous reasoning by properly weighting each question in the evaluation measure. The following is an example of a customized precision measure

$$precision_{method}(x_j) \ = \frac{\sum_{q_i \in Q} w_i \cdot M_j^i}{\sum_{q_i \in Q} w_i \cdot (M_j^i + O_j^i)}, \tag{10}$$

where w_i are the weights assigned by the user to each question.

3 Empirical analysis

We propose an alternate view on the problem of summarization through machine learning formalism, that will allow us to introduce a method of performing controlled experiments to compare document summarizers.

3.1 Machine learning in text summarization

There is a large body of methods to automatically compose summaries [13], and several different ways of grouping these methods (M. Kan and K. McKeown [12], Sparck Jones [11]). However, these groupings are independent from the evaluation techniques, which do not make any distinction between them [10].This makes it difficult to evaluate the specific improvements each summarization method incorporates.

Dissecting the problem into smaller steps has often been useful to reason about the similarities among a priori different methods. Hence we propose a partition of the summarization problem, similar to what is described in [11], which will make comparison of different methods suitable for identifying key implementation choices.

We consider text summarization as a three step procedure

1. Inference or construction of a model of the text,
2. Feature selection, that is selection of units of content,
3. Compilation of the selected units of content.

The inference step characterizes each document in terms of a set of latent variables that may account for syntax, lexical cohesion [7], segmentation [9], and/or statistics [21] of the document. The goal of the second step is to choose the subset of units of content identified in the inference step, that best describes the entire document up to a certain predetermined precision. The last step (compilation) recombines the units of content of the document into a reader-friendly version, and often requires natural language processing (NLP) techniques. Table 1 shows mappings of three known text summarization methods into these three steps.

For each step, the choices of representation, the assumptions, and the goal, define the final summarization algorithm. Therefore, controlled experiments to compare different algorithms with different implementation choices for each step are necessary to rank the key advances each algorithm incorporates.

3.2 Results

Having introduced our categorical evaluation system and a methodology for performing controlled experiments to compare summarization methods, in this section we analyze the differences between two widely used summarization techniques: lead-based summarization and headline based summarization. Despite its simplicity, lead-based summarization has been found to be the best summarizer for news stories [5], and this sole fact justifies studying it and comparing it to its closest relative, headline-based summarization.

Table 1: Examples of known algorithms as described by our dissection of text summarization

	Step	**Representation**	**Assumptions**	**Algorithm**
Lexical Chains summarization [2]	*Inference*	Word sequence	Lexical Cohesion, Distance metric in wordnet [6].	Greedy strategy with word 'memory', lexical chain pruning.
	Feature selection	Sentences, lexical chains	Important sentences are the highest connected entities	Sentence scoring through lexical chains
	Compilation	Sentences	Independence between units of content	Display selection in order of appearance
Lead-Based	*Inference*	None		
	Feature selection	Sentences	Important sentences are located at the beginning of document [4],[5]	Pick the first n sentences
	Compilation	Sentences	Independence between units of content	Display selection in order of appearance
Headline-Based	*Inference*	None		
	Feature selection	Sentences and words	Important sentences contain words used in the headline [4]	Evaluate word co-occurrence with headline and select the n topmost sentences
	Compilation	Sentences	Independence between units of content	Display selection in order of appearance

Data:

Reuters Corpus; Volume 1: English Language, 1996-02-20 to 1997-08-19. Released on: 2000-11-03. Format Version: 1. Correction Level:0. http://about.reuters.com/researchandstandards/corpus/.

We divided the corpus into a Training set ($\mathcal{T}$: October 96, 71708 documents) and two validation sets ($\mathcal{V}_1$: November 96, 67737 documents; $\mathcal{V}_2$: March 97, 64954 documents). The hierarchical categories of the corpus provided the questions for the partial order Q. The first two layers of the hierarchy were used (4 categories in 1^{st} layer and 37 in the 2^{nd}). The hierarchy of classifiers was trained using maximum entropy on a "bag of words" representation [15] of the summaries analyzed. The documents in the Reuters' Corpus come in XML format, and sentences (so called here due to its length and unity of content) come in between paragraph marks (`<p>...</p>`). Stopwords were removed from the documents in all the experiments that follow.

3.2.1 Lead Based vs Headline based

We compare lead based and headline based summarization, as defined in table 1. From the analysis in section 3.1, this comparison shall determine which feature selection method is better in the absence of additional steps.

Headline based

The headline based method for sentence selection assesses the similarity between each sentence of the document and the headline, and selects those sentences ranked above certain threshold of similarity. We consider a measure of similarity between two sentences based on a measure of co-occurrence of words modified from [1], frequently used in document retrieval and document clustering [6]:

$$c(\phi_i, \phi_h) = \frac{\frac{N_{i,h}}{2}}{N_i + N_h - \frac{N_{i,h}}{2}} \tag{11}$$

[6] Other measures of similarity were analyzed, in particular, mutual information was discarded because it required an exponential number of computations to compare the superset of the words used in each sentence and the headline

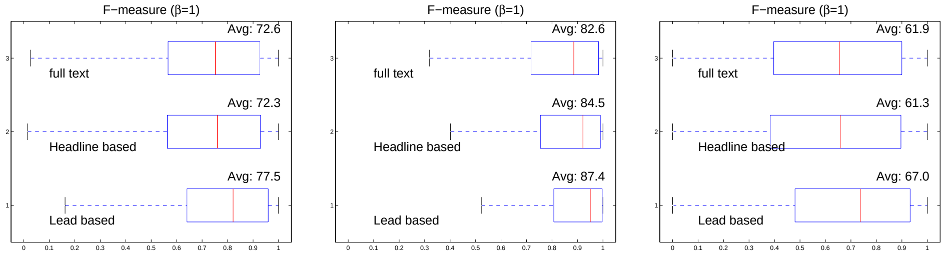
where N_i and N_h stand for the number of words that occur in ϕ_i and ϕ_h respectively. And $N_{i,h}$ is the count, with repetition, of the words that occur in both. The threshold for selection was set according to a cross validation method to an 80% of agreement yielding summaries of 2.5 sentences in average (including the headline), a significant reduction from the average number of sentences in the full document: 11.

Lead based

The lead based method of sentence selection, selects the first N sentences of the document. For the sake of this comparison, N as set to 2,to match the average number of sentences selected by headline based summarization. Additional experiments showed high insensitivity to changing the value of N between 1 and 4, even when N was set specifically for each document to match the number of sentences chosen by headline based summarization for teh same document.

Results

Figure 1 shows the distribution of the F-measure (9) in the $\mathcal{V}_2$ set for lead based, headline based, and the full text representations of the corpus. Additionally, we performed a Kolmogorov Smirnov (K-S) statistical test [17], to do pairwise comparisons of the distributions of the F-measure on the different representations of the documents shown in figure 1. Lead based is significantly different from the rest with $p < 1e - 16$, and headline based and full text are different with $p = 6e - 4$.



(a) (b) (c)

Figure 1: Comparison of the distribution of documents in $\mathcal{V}_2$ for lead based, headline based, and full text representations. (a) On the entire hierarchy, (b) on the first layer, (c) on the second layer.

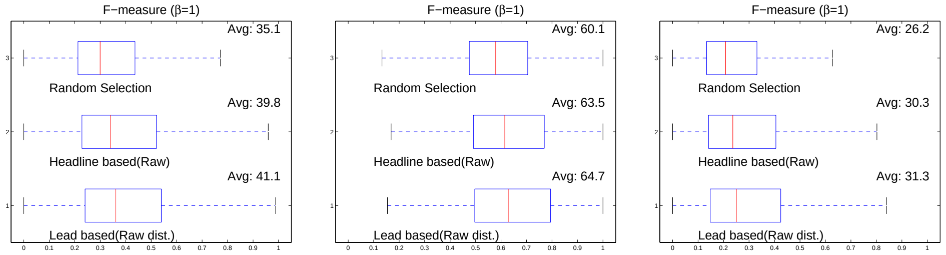
3.2.2 Random Sentence Selection

In order to show the effectiveness of the evaluation method in different conditions, figure 2 compares lead based and headline based summarization against random sentence selection. As training and testing with a random selection of sentences would yield no information about the quality of the summaries, for this experiment, the hierarchy of classifiers was trained on the full text representation of the documents and tested against each of the alternate representations. This explains the low individual performances. Figure 2 shows that both lead and headline based summarizers equally outperform random selection.

3.2.3 An attempt at improving Headline Based

The fact that lead based, a method for sentence selection independent from the content of the document outperforms headline based selection, as shown in the previous sections is intriguing. The following two hypothesis may contribute to explain that result:

- The documents, news stories, are written to stretch the main information in the first lines. The strict *telegraphic* writing style of the news stories seems to support that hypothesis.



(a) (b) (c)

Figure 2: Comparison of the distribution of documents in $\mathcal{V}_2$ for lead based, headline based, and random selection models. The hierarchy was trained on the full text representation. The number of sentences selected for each document and summarization algorithm was given by the number of sentences selected by the headline based criteria on each document.(a) On the entire hierarchy, (b) on the first layer, (c) on the second layer.

- The comparison in the headline based selection algorithm is empoverished by the lack of a consistent model of language. The statistical comparison lacks the ability to relate different words that refer to the same concept.

Looking at the disection of the summarization problem from the beginning of section 3 we introduce a model of language in the inference step of headline based summarization, and thus examine the second hypothesis. Obviously, the lead based criteria for sentence selection would not benefit from such an inference step.

Model of Language

We consider language as a method to express *ideas* ($\mathcal{I}$) in the form of an structured discourse. In this model, words w appear as the only observables (i.e. random instances) of an *idea*:

$$P(w_j|\mathcal{I}_i) = \frac{P(\mathcal{I}_i|w_j)P(w_j)}{P(\mathcal{I}_j)} \tag{12}$$

Each word must represent at least one *idea*

$$\exists\ i\ |\ P(w_j|\mathcal{I}_i) \neq 0 \tag{13}$$

and if more than one, its meaning shall depend on the context surrounding the word. The context must be fully determined by the *ideas* preceding the current one, thus verifying

$$P(I_t) = \sum_{\mathcal{I}_1,\ldots,\mathcal{I}_{t-1}} P(I_t|I_1,\ldots,I_{t-1})P(I_1,\ldots,I_{t-1}) \tag{14}$$

where I_t is the *idea* being analyzed at time t, and $I_1,\ldots,I_{t-1}$ represents a sequence of *ideas* from the distribution of all possible sequences of *ideas* of length $t-1$. Note the use of script letters to represent actual random variables such as *ideas* and non-script letters to represent possible values an idea may take.

Equations (12), (13), (14) describe our abstraction for defining a language.

Meaning Inference

Let us consider a document as a finite sequence of words that results from random sampling with probability

$$P(w_j|\mathcal{I}_t)$$

from the distribution of the corresponding sequence of *ideas*; t reflects the relative order in which the *idea* occurred in the text and w_j is a given word in our vocabulary.
For convenience we assume that context presents the Markov property, so equation (14) becomes:

$$P(I_t) = \sum_{\mathcal{I})} P(I_t|I_{t-1})P(I_{t-1}) \tag{15}$$

equation (15) turns the model for sequences of *ideas* into a Markov model. Since *ideas* are only observed through words, the model of language we propose is a Hidden Markov Model (HMM)[18], and meaning of the words can be inferred through the *maximum a posteriori (MAP) hidden state sequence*

$$P(\mathcal{I}_1...\mathcal{I}_n|w_1,...w_n) \tag{16}$$

or the *sequence of maximum a posteriori hidden states*

$$P(\mathcal{I}_t|w_1,...w_n)\forall t \in 1,..,n \tag{17}$$

where $\mathcal{I}_i$ is the meaning associated with the position t in the text and w_i, the word observed at that position.

The inference of the sequence of *ideas* (the states of our HMM), will couple each word in the document with an *idea*[7]. Replacement of words by *ideas* in the original document will allow to increase the similarities between sentences for the headline based algorithm for sentence selection.

Note that tracking the words coupled with each *idea* in the sequence will yield the thread of an idea, similar to the lexical chains described in [2].

Two options are preferred to train the Hidden Markov model:

1. Estimate a Hidden Markov Model from a set of documents parsed as sequences of "known" words. This may be achieved through the EM algorithm.

2. Learn both the structure of the Hidden Markov Model and its parameters. By either best-first merge [16], entropic learning [3], or applying some form of clustering or latent analysis on the graph defined by a word co-occurrence matrix.

We chose the EM algorithm because, despite its disadvantages (it requires to fix the number of *ideas* $\equiv$ *states*; it is only guaranteed to converge to a local minima; is difficult to scale to large vocabulary sizes), the other two options, namely, structure and parameter learning, require a prohibitive number of calculations for a large vocabulary as that of English Language.

Parameters

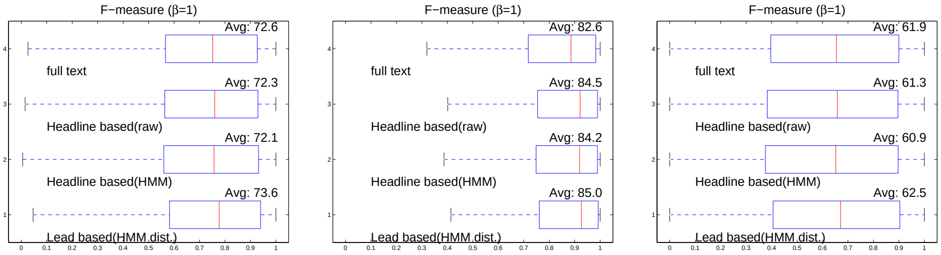
We trained an HMM assuming 20 states (*Ideas*) over a vocabulary including the 90th percentile (1186 words) of our training set once all stopwords had been removed. The EM algorithm was fed with a topology favoring self-transitions and ended when the ratio of change in the loglikelihood went below 10^{-4}.[8]

Results

Figure 3, compares all the representations introduced so far, and shows that the introduction of the hidden markov model in the so called 'augmented' headline based summarizer did not produce any improvement. The explanation for that may be found in the strict "telegraphic" style of Reuters's news.

[7] Words that are unknown to our HMM will not be coupled with an *idea*. This bares some similarity with human understanding of documents: ignoring the meaning of a word does not prevent always from understanding the document.

[8] The choice to use a low number of *ideas* (20) aimed to increase similarity between sentences with lower rates of matching words. The parameters for the HMM were decided after examination of HMMs of 10, 20, 50 and 100 states trained on different initial conditions and different vocabulary sizes (50th, 60th 70th and 80th percentiles), and the choice was made for the HMM that minimized loglikelihood over a validation set.



(a) (b) (c)

Figure 3: Comparison of the distribution of documents in $\mathcal{V}_2$ for lead based, headline based, and augmented headline based selection models. The hierarchy was trained on each summarizer's training set. The number of sentences selected for each document and summarization algorithm was given by the number of sentences selected by the augmented headline based criteria on each document.(a) On the entire hierarchy, (b) on the first layer, (c) on the second layer.

4 Discussion

These results confirm that lead-based summarization outperforms headline based summarization, and its slightly more sophisticated form including a model for language for news articles. This was expected, as lead-based summarization has been shown to outperform complex summarization methods [5] for news stories, and it was already explained then by their particular writing style. The poor performance of the augmented headline based summarization can be explained by the particular writing style of Reuters' news stories, too tailored for a quick read through lead-based analysis (from both computers and human readers). And brings up the possible inadequacy of news stories for comparison of different summarizers. Additionally, these results confirm the common intuition that summaries are better suited for communicating information with a low level of detail than for communicating detailed facts (compare, for example, figures 1(c) and 1(b).

In general, the results presented in section 3 show that it is possible to compare different summarization methods quantitatively, and incorporate a controlled experiment methodology in the comparison. Indeed, the strategy of pushing the human expertise backwards in the comparison procedure, has made possible comparison of summarization methods over larger corpora.

Nonetheless, there are at least three possible critiques to our evaluation approach. The first relates to its inadequacy to evaluate all the details in the compilation step. Although this method does not attempt to evaluate the style of the final summary, it has been recently suggested [22] that human performance in categorization of documents where all structural information has been removed is noticeably high.

The second critique relates to the influence of the performance of the classifiers on the final measure. High performance on the individual classifiers might be required to detect differences between more complex summarization techniques than the ones analyzed here. This is analogous to requiring a high degree of expertise from human evaluators asked to decide among two high quality summaries. In our hierarchy of classifiers, the choice of the set of questions upper bounds the accuracy of the description of the documents; and the development of a two-dimensional hierarchy should allow the performance of the hierarchy become arbitrarily close to that upper bound. As mentioned above, user preferences also play a role in the choice of the best summarizer, and our hierarchy handles them in choice of the triple $\langle length,\ information\ content,\ level\ of\ detail\rangle$.

The third critique emphasizes the inability of this approach to gage highly specific information. The evaluation method presented here is only able to assess how close may two summarization methods agree to classify documents under a possibly broad set of categories deviced by experts. However, this set of categories being common to all the documents makes document specific information fall beyond its analytical capabilities. For the sake of an example consider a classifier trained to detect when a news story references a meeting between any two personalities, (call it meeting detector), thinking of increasing the number of classifiers to account for all possible (or at least the most relevant) meetings between specific personalities is simply absurd. Instead, complementing the meeting detector with an NLP system able to determine who met, and if that information is still available in the summary, remains a more interesting option. Note that, following with the example, the meeting detector is already able to determine if information about the meeting has been lost, thus NLP techniques naturally complement this categorical evaluation system to gain specificity on the analysis of the documents and summaries.

5 Conclusions

We have formalized a task-based summary evaluation system as a model selection problem. Our evaluation technique consists of a hierarchy of classifiers that quantifies differences between summaries by evaluating their content and level of detail. In practice, we shift human expertise to an earlier step of the process allowing processing of larger corpora. Additionally, the description of the summarization problem introduced in section 3.1, combined with our automated task-based evaluation system, allows performing controlled experiments to identify the choices of implementation that make summarization algorithms successful. During the empirical analysis it has been noticed that the analysis of news stories, usual focus for most summarization algorithms, may suffer from its particular writing style, as it tends to give certain advantage to seemingly unstructured methods such as lead-based summarization.

The analysis of classification tasks involving multiple classes and multiply labeled documents poses a challenge for the application of machine learning to information retrieval. Our approach is of interest not only for the evaluation of summarization techniques, where our automated task-based evaluation system confirms results obtained in previous work through human evaluation, but also for more general information retrieval tasks. This is the case of tasks requiring the analysis of documents that belong simultaneously to multiple different categories, eg. categorization of web sites and e-mail classification.

Acknowledgements

The authors would like to thank Gadi Geiger for the long hours spent trying to polish their English expression.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval, 1999.

[2] R. Barzilay and M. Elhadad. Using lexical chains for text summarization, 1997.

[3] M. Brand. Structure discovery in hidden markov models via an entropic prior and parameter extinction, 1997.

[4] H. Edmunson. New methods in automatic abstracting, 1969.

[5] Brandow et al. Automatic condensation of electronic publications by sentence selection., 1995.

[6] C. Fellbaum. Wordnet: An electronic lexical database, 1998.

[7] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.

[8] V. Hatzivassiloglou and K.R. McKeown. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 172–182, Columbus Ohio, June 1993. Association for Computational Linguistics.

[9] Marti Hearst. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9–16, New Mexico State University, Las Cruces, New Mexico, 1994.

[10] H. Jing, R. Barzilay, C. McKeown, and M. Elhadad. Summarization evaluation methods: Experiments and analysis, 1998.

[11] Karen Sparck Jones. What might be in a summary? In Knorz, Krause, and Womser-Hacker, editors, *Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26, Konstanz, DE, 1993. Universitätsverlag Konstanz.

[12] M. Kan and K. McKeown. Domain independent summarization via extracted term types, 1999.

[13] Inderjeet Mani and Mark T. Maybury, editors. *Advances in Automatic Text Summarization*. MIT Press, July 1999.

[14] Daniel Marcu. From discourse structures to text summaries. In I. Mani and M. Maybury, editors, *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, 1997. UNED.

[15] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification, 1999.

[16] Stephen M. Omohundro. Best-first model merging for dynamic learning and recognition. In John E. Moody, Steve J. Hanson, and Richard P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 958–965. Morgan Kaufmann Publishers, Inc., 1992.

[17] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flanney. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1995. p 623 - 628.

[18] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *IEEE Speech Recognition*, volume 2, pages 257–285. IEEE, 1989.

[19] C.J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 2nd edition, June 1981.

[20] Paul Viola and Michael Jones. Robust real-time object detection, 2001.

[21] Michael J. Witbrock and Vibhu O. Mittal. Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries (poster abstract). In *Research and Development in Information Retrieval*, pages 315–316, 1999.

[22] F. Wolf and P. sinha. Human document classification. Talk at MIT, NSF-ITR meeting.