A Bayesian Method to Incorporate Background Knowledge during Automatic Text Summarization

Annie Louis
ILCC, School of Informatics,
University of Edinburgh,
Edinburgh EH8 9AB, UK
alouis@inf.ed.ac.uk

Abstract

In order to summarize a document, it is often useful to have a *background* set of documents from the domain to serve as a reference for determining new and important information in the input document. We present a model based on Bayesian surprise which provides an intuitive way to identify surprising information from a summarization input with respect to a background corpus. Specifically, the method quantifies the degree to which pieces of information in the input change one's beliefs' about the world represented in the background. We develop systems for generic and update summarization based on this idea. Our method provides competitive content selection performance with particular advantages in the update task where systems are given a small and topical background corpus.

1 Introduction

Important facts in a new text are those which deviate from previous knowledge on the topic. When people create summaries, they use their knowledge about the world to decide what content in an input document is informative to include in a summary. Understandably in automatic summarization as well, it is useful to keep a background set of documents to represent general facts and their frequency in the domain.

For example, in the simplest setting of multi-document summarization of news, systems are asked to summarize an *input set* of topically-related news documents to reflect its central content. In this GENERIC task, some of the best reported results were obtained by a system (Conroy et al., 2006) which computed importance scores for words in the input by examining if the word occurs with significantly higher probability in the input compared to a large background collection of news articles. Other specialized summarization tasks explicitly require the use of background information. In the UPDATE summarization task, a system is given two sets of news documents on the same topic; the second contains articles published later in time. The system should summarize the important updates from the second set assuming a user has already read the first set of articles.

In this work, we present a Bayesian model for assessing the novelty of a sentence taken from a summarization input with respect to a background corpus of documents.

Our model is based on the idea of Bayesian Surprise (Itti and Baldi, 2006). For illustration, assume that a user's background knowledge comprises of multiple hypotheses about the current state of the world and a probability distribution over these hypotheses indicates his degree of belief in each hypothesis. For example, one hypothesis may be that *the political situation in Ukraine is peaceful*, another where *it is not*. Apriori assume the user favors the hypothesis about a peaceful Ukraine, i.e. the hypothesis has higher probability in the prior distribution. Given new data, the evidence can be incorporated using Bayes Rule to compute the posterior distribution over the hypotheses. For example, upon viewing news reports about riots in the country, a user would update his beliefs and the posterior distribution of the user's knowledge would have a higher probability for a riotous Ukraine. Bayesian surprise is the difference between the prior and posterior distributions over the hypotheses which quantifies the extent to which the new data (the news report) has changed a user's prior beliefs about the world.

In this work, we exemplify how Bayesian surprise can be used to do content selection for text summarization. Here a user's prior knowledge is approximated by a background corpus and we

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 333–338,
Baltimore, Maryland, USA, June 23-25 2014.

show how to identify sentences from the input set which are most surprising with respect to this background. We use the method to do two types of summarization tasks: a) GENERIC news summarization which uses a large random collection of news articles as the background, and b) UPDATE summarization where the background is a smaller but specific set of news documents on the same topic as the input set. We find that our method performs competitively with a previous log-likelihood ratio approach which identifies words with significantly higher probability in the input compared to the background. The Bayesian approach is more advantageous in the update task, where the background corpus is smaller in size.

2 Related work

Computing new information is useful in many applications. The TREC novelty tasks (Allan et al., 2003; Soboroff and Harman, 2005; Schiffman, 2005) tested the ability of systems to find novel information in an IR setting. Systems were given a list of documents ranked according to relevance to a query. The goal is to find sentences in each document which are relevant to the query, and at the same time is new information given the content of documents higher in the relevance list.

For update summarization of news, methods range from textual entailment techniques (Bentivogli et al., 2010) to find facts in the input which are not entailed by the background, to Bayesian topic models (Delort and Alfonseca, 2012) which aim to learn and use topics discussed only in background, those only in the update input and those that overlap across the two sets.

Even for generic summarization, some of the best results were obtained by Conroy et al. (2006) by using a large random corpus of news articles as the background while summarizing a new article, an idea first proposed by Lin and Hovy (2000). Central to this approach is the use of a likelihood ratio test to compute *topic words*, words that have significantly higher probability in the input compared to the background corpus, and are hence descriptive of the input's topic. In this work, we compare our system to topic word based ones since the latter is also a general method to find surprising new words in a set of input documents but is not a bayesian approach. We briefly explain the topic words based approach below.

Computing topic words: Let us call the input set I and the background B. The log-likelihood ratio test compares two hypotheses:

H_1: A word t is not a topic word and occurs with equal probability in I and B, i.e. $p(t|I) = p(t|B) = p$

H_2: t is a topic word, hence $p(t|I) = p_1$ and $p(t|B) = p_2$ and $p_1 > p_2$

A set of documents D containing N tokens is viewed as a sequence of words $w_1 w_2 ... w_N$. The word in each position i is assumed to be generated by a Bernoulli trial which succeeds when the generated word $w_i = t$ and fails when w_i is not t. Suppose that the probability of success is p. Then the probability of a word t appearing k times in a dataset of N tokens is the binomial probability:

$$b(k, N, p) = \binom{N}{k} p^k (1-p)^{N-k} \quad (1)$$

The likelihood ratio compares the likelihood of the data $D = \{B, I\}$ under the two hypotheses.

$$\lambda = \frac{P(D|H_1)}{P(D|H_2)} = \frac{b(c_t, N, p)}{b(c_I, N_I, p_1)\, b(c_B, N_B, p_2)} \quad (2)$$

p, p_1 and p_2 are estimated by maximum likelihood. $p = c_t/N$ where c_t is the number of times word t appears in the total set of tokens comprising $\{B, I\}$. $p_1 = c_t^I/N_I$ and $p_2 = c_t^B/N_B$ are the probabilities of t estimated only from the input and only from the background respectively.

A convenient aspect of this approach is that $-2 \log \lambda$ is asymptotically χ^2 distributed. So for a resulting $-2 \log \lambda$ value, we can use the χ^2 table to find the significance level with which the null hypothesis H_1 can be rejected. For example, a value of 10 corresponds to a significance level of 0.001 and is standardly used as the cutoff. Words with $-2 \log \lambda > 10$ are considered topic words. Conroy et al. (2006)'s system gives a weight of 1 to the topic words and scores sentences using the number of topic words normalized by sentence length.

3 Bayesian Surprise

First we present the formal definition of Bayesian surprise given by Itti and Baldi (2006) without reference to the summarization task.

Let **H** be the space of all hypotheses representing the background knowledge of a user. The user has a probability $P(H)$ associated with each hypothesis $H \in \mathbf{H}$. Let D be a new observation. The posterior probability of a single hypothesis H can be computed as:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \quad (3)$$

The surprise $S(D, \mathbf{H})$ created by D on hypothesis space $\mathbf{H}$ is defined as the difference between the prior and posterior distributions over the hypotheses, and is computed using KL divergence.

$$S(D, \mathbf{H}) = \text{KL}(P(H|D), P(H))) \quad (4)$$

$$= \int_{\mathbf{H}} P(H|D) \log \frac{P(H|D)}{P(H)} \quad (5)$$

Note that since KL-divergence is not symmetric, we could also compute $\text{KL}(P(H), P(H|D))$ as the surprise value. In some cases, surprise can be computed analytically, in particular when the prior distribution is conjugate to the form of the hypothesis, and so the posterior has the same functional form as the prior. (See Baldi and Itti (2010) for the surprise computation for different families of probability distributions).

4 Summarization with Bayesian Surprise

We consider the hypothesis space $\mathbf{H}$ as the set of all the hypotheses encoding background knowledge. A single hypothesis about the background takes the form of a multinomial distribution over word unigrams. For example, one multinomial may have higher word probabilities for 'Ukraine' and 'peaceful' and another multinomial has higher probabilities for 'Ukraine' and 'riots'. $P(H)$ gives a prior probability to each hypothesis based on the information in the background corpus. In our case, $P(H)$ is a Dirichlet distribution, the conjugate prior for multinomials. Suppose that the vocabulary size of the background corpus is V and we label the word types as (w_1, w_2, ... w_V). Then,

$$P(H) = Dir(\alpha_1, \alpha_2, ...\alpha_V) \quad (6)$$

where $\alpha_{1:V}$ are the concentration parameters of the Dirichlet distribution (and will be set using the background corpus as explained in Section 4.2).

Now consider a new observation I (a text, sentence, or paragraph from the *summarization input*) and the word counts in I given by (c_1, c_2, ..., c_V). Then the posterior over H is the dirichlet:

$$P(H|I) = Dir(\alpha_1 + c_1, \alpha_2 + c_2, ..., \alpha_V + c_V) \quad (7)$$

The surprise due to observing I, $S(I, \mathbf{H})$ is the KL divergence between the two dirichlet distributions. (Details about computing KL divergence between two dirichlet distributions can be found in Penny (2001) and Baldi and Itti (2010)).

Below we propose a general algorithm for summarization using surprise computation. Then we define the prior distribution $P(H)$ for each of our two tasks, GENERIC and UPDATE summarization.

4.1 Extractive summarization algorithm

We first compute a surprise value for each word type in the summarization input. Word scores are aggregated to obtain a score for each sentence.

Step 1: Word score. Suppose that word type w_i appears c_i times in the summarization input I. We obtain the posterior distribution after seeing all instances of this word ($\mathbf{w_i}$) as $P(H|\mathbf{w_i}) = Dir(\alpha_1, \alpha_2, ...\alpha_i + c_i, ...\alpha_V)$. The score for w_i is the surprise computed as KL divergence between $P(H|\mathbf{w_i})$ and the prior $P(H)$ (eqn. 6).

Step 2: Sentence score. The composition functions to obtain sentence scores from word scores can impact content selection performance (Nenkova et al., 2006). We experiment with sum and average value of the word scores.[1]

Step 3: Sentence selection. The goal is to select a subset of sentences with high surprise values. We follow a greedy approach to optimize the summary surprise by choosing the most surprising sentence, the next most surprising and so on. At the same time, we aim to avoid redundancy, i.e. selecting sentences with similar content. After a sentence is selected for the summary, the surprise for words from this sentence are set to zero. We recompute the surprise for the remaining sentences using step 2 and the selection process continues until the summary length limit is reached.

The key differences between our Bayesian approach and a method such as topic words are: (i) The Bayesian approach keeps multiple hypotheses about the background rather than a single one. Surprise is computed based on the changes in probabilities of all of these hypotheses upon seeing the summarization input. (ii) The computation of topic words is local, it assumes a binomial distribution and the occurrence of a word is independent of others. In contrast, word surprise although computed for each word type separately, quantifies the surprise when incorporating the new counts of this word into the background multinomials.

4.2 Input and background

Here we describe the input sets and background corpus used for the two summarization tasks and

[1] An alternative algorithm could directly compute the surprise of a sentence by incorporating the words from the sentence into the posterior. However, we found this specific method to not work well probably because the few and unrepeated content words from a sentence did not change the posterior much. In future, we plan to use latent topic models to assign a topic to a sentence so that the counts of all the sentence's words can be aggregated into one dimension.

define the prior distribution for each. We use data from the DUC[2] and TAC[3] summarization evaluation workshops conducted by NIST.

Generic summarization. We use multidocument inputs from DUC 2004. There were 50 inputs, each contains around 10 documents on a common topic. Each input is also provided with 4 manually written summaries created by NIST assessors. We use these manual summaries for evaluation.

The background corpus is a collection of 5000 randomly selected articles from the English Gigaword corpus. We use a list of 571 stop words from the SMART IR system (Buckley, 1985) and the remaining content word vocabulary has 59,497 word types. The count of each word in the background is calculated and used as the α parameters of the prior Dirichlet distribution $P(H)$ (eqn. 6).

Update summarization. This task uses data from TAC 2009. An input has two sets of documents, A and B, each containing 10 documents. Both A and B are on same topic but documents in B were published at a later time than A (background). There were 44 inputs and 4 manual update summaries are provided for each.

The prior parameters are the counts of words in A for that input (using the same stoplist). The vocabulary of these A sets is smaller, ranging from 400 to 3000 words for the different inputs.

In practice for both tasks, a new summarization input can have words unseen in the background. So *new* words in an input are added to the background corpus with a count of 1 and the counts of *existing* words in the background are incremented by 1 before computing the prior parameters. The summary length limit is 100 words in both tasks.

5 Systems for comparison

We compare against three types of systems, (i) those which similarly to surprise, use a background corpus to identify important sentences, (ii) a system that uses information from the input set only and no background, and (iii) systems that combine scores from the input and background.

KL_{back}: represents a simple baseline for surprise computation from a background corpus. A *single* unigram probability distribution B is created from the background using maximum likelihood. The summary is created by greedily adding sentences which maximize KL divergence between B and the current summary. Suppose the set of sentences currently chosen in the summary is S. The next step chooses the sentence $s_l = \arg\max_{s_i} \text{KL}(\{S \cup s_i\}||B)$.

TS_{sum}, TS_{avg}: use topic words computed as described in Section 2 and utilizing the same background corpus for the generic and update tasks as the surprise-based methods. For the generic task, we use a critical value of 10 (0.001 significance level) for the χ^2 distribution during topic word computation. In the update task however, the background corpus A is smaller and for most inputs, no words exceeded this cutoff. We lower the significance level to the generally accepted value of 0.05 and take words scoring above this as topic words. The number of topic words is still small (ranging from 1 to 30) for different inputs.

The TS_{sum} system selects sentences with greater counts of topic words and TS_{avg} computes the number of topic words normalized by sentence length. A greedy selection procedure is used. To reduce redundancy, once a sentence is added, the topic words contained in it are removed from the topic word list before the next sentence selection.

KL_{inp}: represents the system that *does not use* background information. Rather the method creates a summary by optimizing for high similarity of the summary with the input word distribution.

Suppose the input unigram distribution is I and the current summary is S, the method chooses the sentence $s_l = \arg\min_{s_i} \text{KL}(\{S \cup s_i\}||I)$ at each iteration. Since $\{S \cup s_i\}$ is used to compute divergence, redundancy is implicitly controlled in this approach. Such a KL objective was used in competitive systems in the past (Daumé III and Marcu, 2006; Haghighi and Vanderwende, 2009).

Input + background: These systems combine (i) a score based on the background (KL_{back}, TS or SR) with (ii) the score based on the input only (KL_{inp}). For example, to combine TS_{sum} and KL_{inp}: for each sentence, we compute its scores based on the two methods. Then we normalize the two sets of scores for candidate sentences using z-scores and compute the best sentence as $\arg\max_{s_i}$ ($\text{TS}_{\text{sum}}(s_i)$ - $\text{KL}_{\text{inp}}(s_i)$). Redundancy control is done similarly to the TS only systems.

6 Content selection results

For evaluation, we compare each summary to the four manual summaries using ROUGE (Lin and Hovy, 2003; Lin, 2004). All summaries were truncated to 100 words, stemming was performed and

[2] http://www-nlpir.nist.gov/projects/duc/index.html

[3] http://www.nist.gov/tac/

	ROUGE-1	ROUGE-2
KL_{back}	0.2276 (TS, SR)	0.0250 (TS, SR)
TS_{sum}	0.3078	**0.0616**
TS_{avg}	0.2841 (TS_{sum}, SR_{sum})	0.0493 (TS_{sum})
SR_{sum}	**0.3120**	0.0580
SR_{avg}	0.3003	0.0549
KL_{inp}	0.3075 (KL_{inp}+TS_{avg})	0.0684
KL_{inp}+TS_{sum}	0.3250	0.0725
KL_{inp}+TS_{avg}	**0.3410**	**0.0795**
KL_{inp}+SR_{sum}	0.3187 (KL_{inp}+TS_{avg})	0.0660 (KL_{inp}+TS_{avg})
KL_{inp}+SR_{avg}	0.3220 (KL_{inp}+TS_{avg})	0.0696

Table 1: Evaluation results for generic summaries. Systems in parentheses are significantly better.

	ROUGE-1	ROUGE-2
KL_{back}	0.2246 (TS, SR)	0.0213 (TS, SR)
TS_{sum}	0.3037 (SR_{avg})	0.0563
TS_{avg}	0.2909 (SR_{sum}, SR_{avg})	0.0477 (SR_{sum}, SR_{avg})
SR_{sum}	0.3201	**0.0640**
SR_{avg}	**0.3226**	0.0639
KL_{inp}	0.3098 (KL_{inp}+SR_{avg})	0.0710
KL_{inp}+TS_{sum}	0.3010 (KL_{inp}+$SR_{sum, avg}$)	0.0635
KL_{inp}+TS_{avg}	0.3021 (KL_{inp}+$SR_{sum, avg}$)	0.0543 (KL_{inp}, KL_{inp}+$SR_{sum, avg}$)
KL_{inp}+SR_{sum}	0.3292	0.0721
KL_{inp}+SR_{avg}	**0.3379**	**0.0767**

Table 2: Evaluation results for update summaries. Systems in parentheses are significantly better.

stop words were **not** removed, as is standard in TAC evaluations. We report the ROUGE-1 and ROUGE-2 recall scores (average over the inputs) for each system. We use the Wilcoxon signed-rank test to check for significant differences in mean scores. Table 1 shows the scores for generic summaries and 2 for the update task. For each system, the peer systems with significantly better scores (p-value < 0.05) are indicated within parentheses.

We refer to the surprise-based summaries as $\mathbf{SR_{sum}}$ and $\mathbf{SR_{avg}}$ depending on the type of composition function (Section 4.1).

First, consider GENERIC summarization and the systems which use the background corpus only (those above the horizontal line). The KL_{back} baseline performs significantly worse than topic words and surprise summaries. Numerically, SR_{sum} has the highest ROUGE-1 score and TS_{sum} tops according to ROUGE-2. As per the Wilcoxon test, TS_{sum}, SR_{sum} and SR_{avg} scores are statistically indistinguishable at 95% confidence level.

Systems below the horizontal line in Table 1 use an objective which combines both similarity with the input and difference from the background. The first line here shows that a system optimizing only for input similarity, KL_{inp}, by itself has higher scores (though not significant) than those using background information only. This result is not surprising for generic summarization where all the topical content is present in the input and the background is a non-focused random collection. At the same time, adding either TS or SR scores to KL_{inp} almost always leads to better results with $KL_{inp} + TS_{avg}$ giving the best score.

In UPDATE summarization, the surprise-based methods have an advantage over the topic word ones. SR_{avg} is significantly better than TS_{avg} for both ROUGE-1 and ROUGE-2 scores and better than TS_{sum} according to ROUGE-1. In fact, the surprise methods have numerically higher ROUGE-1 scores compared to input similarity (KL_{inp}) in contrast to generic summarization. When combined with KL_{inp}, the surprise methods provide improved results, significantly better in terms of ROUGE-1 scores. The TS methods do not lead to any improvement, and $KL_{inp} + TS_{avg}$ is significantly worse than KL_{inp} only. The limitation of the TS approach arises from the paucity of topic words that exceed the significance cutoff applied on the log-likelihood ratio. But Bayesian surprise is robust on the small background corpus and does not need any tuning for cutoff values depending on the size of the background set.

Note that these models do not perform on par with summarization systems that use multiple indicators of content importance, involve supervised training and which perform sentence compression. Rather our goal in this work is to demonstrate a simple and intuitive unsupervised model.

7 Conclusion

We have introduced a Bayesian summarization method that strongly aligns with intuitions about how people use existing knowledge to identify important events or content in new observations.

Our method is especially valuable when a system must utilize a small background corpus. While the update task datasets we have used were carefully selected and grouped by NIST assesors into initial and background sets, for systems on the web, there is little control over the number of background documents on a particular topic. A system should be able to use smaller amounts of background information and as new data arrives, be able to incorporate the evidence. Our Bayesian approach is a natural fit in such a setting.

Acknowledgements

The author was supported by a Newton International Fellowship (NF120479) from the Royal Society and the British Academy.

References

J. Allan, C. Wade, and A. Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *Proceedings of SIGIR*, pages 314–321.

P. Baldi and L. Itti. 2010. Of bits and wows: a bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666.

L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo. 2010. The sixth PASCAL recognizing textual entailment challenge. *Proceedings of TAC*.

C. Buckley. 1985. Implementation of the SMART information retrieval system. Technical report, Cornell University.

J. Conroy, J. Schlesinger, and D. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of COLING-ACL*, pages 152–159.

H. Daumé III and D. Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of ACL*, pages 305–312.

J. Delort and E. Alfonseca. 2012. DualSum: A topic-model based approach for update summarization. In *Proceedings of EACL*, pages 214–223.

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of NAACL-HLT*, pages 362–370.

L. Itti and P. F. Baldi. 2006. Bayesian surprise attracts human attention. In *Proceedings of NIPS*, pages 547–554.

C. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, pages 495–501.

C. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, pages 1085–1090.

C. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Text Summarization Branches Out Workshop, ACL*, pages 74–81.

A. Nenkova, L. Vanderwende, and K. McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of SIGIR*, pages 573–580.

W. D Penny. 2001. Kullback-liebler divergences of normal, gamma, dirichlet and wishart densities. *Wellcome Department of Cognitive Neurology*.

B. Schiffman. 2005. *Learning to Identify New Information*. Ph.D. thesis, Columbia University.

I. Soboroff and D. Harman. 2005. Novelty detection: the trec experience. In *Proceedings of HLT-EMNLP*, pages 105–112.