Document Summarization Based on Word Associations

Oskar Gross
Department of Computer Science and HIIT
University of Helsinki, Finland
oskar.gross@cs.helsinki.fi

Antoine Doucet
GREYC, CNRS UMR 6072
University of Normandy, Unicaen, France
antoine.doucet@unicaen.fr

Hannu Toivonen
Department of Computer Science and HIIT
University of Helsinki, Finland
hannu.toivonen@cs.helsinki.fi

ABSTRACT

In the age of big data, automatic methods for creating summaries of documents become increasingly important. In this paper we propose a novel, unsupervised method for (multi-)document summarization. In an unsupervised and language-independent fashion, this approach relies on the strength of word associations in the set of documents to be summarized. The summaries are generated by picking sentences which cover the most specific word associations of the document(s). We measure the performance on the DUC 2007 dataset. Our experiments indicate that the proposed method is the best-performing unsupervised summarization method in the state-of-the-art that makes no use of human-curated knowledge bases.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language models—*abstracting methods, summarization*

General Terms

Algorithms, Experimentation, Languages

Keywords

Multi-Document Summarization, Word Associations

1. INTRODUCTION

We propose a novel method for document summarization, Association Mixture Text Summarization, aimed to abstract a news story into a shorter text. Like most other methods, Association Mixture Text Summarization works in a sentence-based manner, selecting a set of sentences from the document to be summarized to constitute its summary. The sentences are chosen so that they collectively cover as much of the relevant information in the original document as possible. The main difficulties are to define what is relevant and to measure how well sets of sentences cover relevant information. Our method has three central characteristics:

(1) *Relevance is based on the relative associations between words, helping to grasp the most salient information in a news story.* Much of the core content of news stories is in the links they establish, e.g., between people, acts, events, and places. We argue that associations at subtler levels can also be important, even ones between adjectives or adverbs and noun or verbs used in the news. Recognition of associations is based on statistical analysis of word co-occurrences within sentences. We believe that such associations reflect the key ideas of news and are useful for selecting sentences.

(2) *Novel associations in a document are recognized by contrasting them against a background corpus.* News stories are supposed to tell something new and a key problem in summarization is to identify what is new in a given document. We treat this as a novelty detection task by contrasting the document to a background corpus to see which associations are emphasized more in the document.

(3) *Natural language processing is trivial, making the method language-independent.* All processed documents are split to sentences and tokens (words) based on punctuation and whitespaces; numbers are removed, and the remaining tokens are used as they are, without any further processing.

In this paper we focus on the sentence selection subtask of document summarization. We do not address the issue of arranging or processing the sentences for improved readability. We evaluate the method in English using public benchmarks, and leave experiments with other languages for future work. In the experiments, our proposed method outperforms all unsupervised summarization methods that do not use semantic resources such as Wordnet.

This paper is organised as follows. We next briefly review related work. We then present the Association Mixture Text Summarization method in Section 3. The performance of the method is evaluated in Section 4, while Section 5 concludes this article with a discussion.

2. RELATED WORK

Document summarization is a well-studied area. There are two types of summarizations methods: methods which select existing sentences and methods which generate sentences. Both of these types of methods can be either supervised or unsupervised, i.e., either learning from examples of existing summaries or not. We focus on the unsupervised domain, of which we give a very brief overview. Nenkova and McKeown [10] provide an exhaustive review of the topic.

Some methods use Latent Semantic Analysis (LSA) [2] as their basis (e.g. [4]). The state-of-the art in purely unsupervised summarization is represented by the DSDR method of

SIGIR '14, July 06 - 11 2014, Gold Coast , QLD, Australia
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609500

He et al. [5]. This approach generates a summary by using sentences that best "reconstruct" the original document. This work has been extended by Zhang et al. [11] who combined document reconstruction and topic decomposition.

A number of unsupervised methods take advantage of additional linguistic resources. In particular, the Two-Tiered Topic model by Celikyilmaz [1] uses Wordnet [9] and the DUC-provided user query for selecting the summary sentences. The Document Understanding Conference[1] (DUC) provides most evaluation procedures and collections in the summarization field. We provide further details in Section 4.

3. METHOD

The *Association Mixture Text Summarization* method proposed below takes as its input a *document* D to be summarized and a *background corpus* $\mathcal{B}$ consisting of a set of documents representing the norm or the current state of information.

As a special case, the background corpus can be empty. Additionally, by extension, instead of a single document a set of documents can be summarized by simply giving their concatenation as the input document D, as will be done in the experimental section.

The method has two parts: (1) computation of document-specific word associations, and (2) selection of sentences with strong word associations. These two steps are described in the following subsections.

3.1 Finding Document-Specific Associations

We consider two relevance criteria for associations in the given document D.

First, an association between two words is more relevant if they are statistically mutually dependent, i.e., if they co-occur in D more frequently than they would by chance. This, of course, is a classic idea.

Second, and more interestingly, the association is characteristic for document D if the two words co-occur in D more frequently than in the background corpus $\mathcal{B}$.

The second criterion is in principle more useful since it uses additional data to assess the association, but it is of little value if the background corpus is small or if the words or the word pair does not occur in the corpus. Our method therefore uses a mixture model of the two criteria above.

Notation. We first define the notation for various counts of words and word pairs in document D and in background $\mathcal{B}$. Let t_i and t_j be words. We use n_{ij} to denote the number of sentences in document D that contain both words t_i and t_j, n_{i-j} the number of sentences containing word t_i but not t_j, n_{-ij} the number of sentences containing t_j but not t_i, and n_{-i-j} the number of sentences containing neither t_i nor t_j. We use $n_{i\cdot} = n_{ij} + n_{i-j}$ to denote the total number of sentences containg word t_i, and respectively for $n_{\cdot j}$. Let $n = |D|$ denote the total number of sentences in document D. Finally, let m_{ij}, m_{i-j}, m_{-ij}, m_{-i-j}, $m_{i\cdot}$, $m_{\cdot j}$ and m be the respective counts in the background corpus $\mathcal{B}$.

Statistical Model. Consider the association between words t_i and t_j. We use multinomial distributions to model the probabilities of observing different combinations of existence/non-existence of words t_i and t_j in a sentence. The four respective model parameters are p_{ij}, p_{i-j}, p_{-ij} and p_{-i-j}, affecting the likelihood of the observed counts n_{ij}, n_{i-j}, n_{-ij} and n_{-i-j}. Three such models are given next, and the fit of the data to these models is later used to assign a weight to the association between t_i and t_j. The third model is the Association Mixture model, while the first two are simpler models that will be used as the components of the mixture.

For convenience, we below define the models using parameters $p_{i\cdot}$ (the probability of observing word t_i), $p_{\cdot j}$ (the probability of observing word t_j), and p_{ij} (the probability of observing both t_i and t_j). These give more natural definitions for the models. The multinomial model parameters can then easily be obtained as $p_{i-j} = p_{i\cdot} - p_{ij}$; $p_{-ij} = p_{\cdot j} - p_{ij}$; $p_{-i-j} = 1 - p_{ij} - p_{i-j} - p_{-ij}$.

The independence model (component) $p^{\text{D-ind}}$ considers observed frequencies of words t_1 and t_2 only in document D and assumes that they are statistically independent:

$$p_{i\cdot}^{\text{D-ind}} = n_{i\cdot}/n; \quad p_{\cdot j}^{\text{D-ind}} = n_{\cdot j}/n; \quad p_{ij}^{\text{D-ind}} = n_{i\cdot} \cdot n_{\cdot j}/n^2.$$

If the data fits this model badly, i.e., essentially if n_{ij} deviates a lot from $n_{i\cdot} \cdot n_{\cdot j}/n$, then the words are likely to be statistically dependent.

The background model (component) p^{B} estimates all three parameters from the respective relative frequencies in the background corpus $\mathcal{B}$:

$$p_{i\cdot}^{\text{B}} = m_{i\cdot}/m; \quad p_{\cdot j}^{\text{B}} = m_{\cdot j}/m; \quad p_{ij}^{\text{B}} = m_{ij}/m.$$

If the data fits this model badly then the word pair occurs in the document differently from the background. This signals that the association is novel.

The association mixture model $p^{\text{B+D-ind}}$ averages the two components above, weighted by their sample sizes n and m: $p^{\text{B+D-ind}} = (n \cdot p^{\text{D-ind}} + m \cdot p^{\text{B}})/(n+m)$. This gives

$$\begin{aligned} p_{i\cdot}^{\text{B+D-ind}} &= (n_{i\cdot} + m_{i\cdot})/(n+m), \\ p_{\cdot j}^{\text{B+D-ind}} &= (n_{\cdot j} + m_{\cdot j})/(n+m), \\ p_{ij}^{\text{B+D-ind}} &= (n_{i\cdot} \cdot n_{\cdot j}/n + m_{ij})/(n+m). \end{aligned}$$

In other words, the mixture model combines information from document D itself and from the background $\mathcal{B}$. Their relative weights adapt to their relative sizes, giving more emphasis to the statistically more reliable source of information.

Association Weights. The weight of the association between two words is based on a log-likelihood ratio test [3]. The test compares two models for each word pair: (1) a null model, in our case the mixture model, and (2) a maximum likelihood alternative model. If the likelihood of the alternative model is much higher, then the null model is less likely to be true. In other words, the mixture model is an expression of expectations, and we are actually interested in finding exceptions to them.

The maximum likelihood model p^{D} is obtained by simply assigning the model parameters directly from the observed relative frequencies: $p_{i\cdot}^{\text{D}} = n_{i\cdot}/n$; $p_{\cdot j}^{\text{D}} = n_{\cdot j}/n$; $p_{ij}^{\text{D}} = n_{ij}/n$.

Let $L(p^{\text{D}})$ be the likelihood of the maximum likelihood model given the counts n_{ij}, n_{i-j}, n_{-ij}, n_{-i-j} in document D, and let $L(p^{\text{B+D-ind}})$ be the likelihood of the mixture model given the same counts. We define the weight $w(t_i, t_j)$ of the association between t_i and t_j as the value of the respective log-likelihood ratio test:

$$w(t_i, t_j) = -2 \log \frac{L(p^{\text{B+D-ind}})}{L(p^{\text{D}})}.$$

[1] http://duc.nist.gov/

Multinomial coefficients in the likelihoods cancel out, and after simplification we have

$$w(t_i, t_j) = 2 \sum_{\substack{a \in \{\text{“}ij\text{”}, \text{“}i{-}j\text{”},\\ \text{“}{-}ij\text{”}, \text{“}{-}i{-}j\text{”}\}}} n_a(\log p_a^{\mathrm{D}} - \log p_a^{\mathrm{B+D\text{-}ind}}).$$

The log-likelihood ratio test gives lower weights for word pairs that better match the mixture model and higher weights for those associations that are unexpected with respect to the mixture model. In text summarization, we are interested in word pairs that have a higher relative frequency in the document D than in the background $\mathcal{B}$, and that have a high log-likelihood ratio.

3.2 Sentence Selection

The other subtask is to select from document D sentences that contain strong word associations. In the sentence selection phase, our goal is to preserve as many of the stronger associations and thereby as much as possible of the core contents of the original document D.

Given a fixed target size of the summary (e.g. 250 words) and the association weights, we aim to pick sentences such that the sum of the log-likelihood ratios of word pairs in the summary is maximized. To avoid selecting sentences with too similar content, each pair is taken into account once.

Formally, let document D be a set of sentences and let each sentence be a set of words. We call any subset $S = \{s'_1, \ldots, s'_s\} \subset D$ of sentences a *summary* of D. We define the total weight of associations in summary S as

$$w(S) = \sum_{\substack{\{t_i,t_j\} \text{ s.t. } t_i \neq t_j \wedge \\ \exists s \in S\colon \{t_i,t_j\} \subset s}} w(t_i, t_j),$$

i.e., as a sum over the set of word pairs in any sentence of the summary. Every pair is only counted once.

In the sentence selection step we aim to find a summary $S^* \subset D$ with a maximal total weight, i.e.,

$$S^* = \arg \max_{\substack{S \subset D \\ ||S|| \leq L}} w(S),$$

where $||S||$ is the number of words in summary S. In our experiments below, the upper limit is set to $L = 250$ words.

This problem is similar to the weighted set cover problem [6]: use sentences of the document to cover as much of the associations as possible. Due to the limited length of the summary, a natural “cost” of a sentence is the number of words in it. Given the computational complexity of the task, we resort to a greedy algorithm [6] to find a summary S that approximates the optimum S^*.

For the sake of simplicity, in the experiments below we add sentences to the summary S until the maximum size is reached ($||S|| \geq L$) and then simply truncate the summary to L words.

4. EXPERIMENTS

In this section, we describe experiments carried out to evaluate the proposed Association Mixture Text Summarization method. We aim to address the following questions: (1) How does the method perform in comparison to state-of-the-art unsupervised summarization methods? (2) What are the contributions of the components p^{B} and $p^{\mathrm{D\text{-}ind}}$ to the method? (3) What is the effect of the size of the background corpus $\mathcal{B}$ on the quality of the summaries?

4.1 Experimental Setup

For experiments and comparisons we use the DUC 2007 dataset consisting of 45 topics. Each topic of 25 documents from the AQUAINT corpus of English news is to be summarized into a collective abstract of at most 250 words.

The evaluation measure is the well-known ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [8]. We use the model summaries of the DUC datasets and their associated tools to compute the ROUGE measures. According to Lin and Hovy [7] the ROUGE-1 score has the best correspondence with human judgements. It is therefore the main focus of our evaluation. We experimented with several background corpora: the Brown corpus, the Gütenberg corpus, the Reuters RCV-1 corpus, as well as combinations.

Data Preprocessing. We remove all markup tags from the documents and leave only the headline and textual content of the news story. We then split the content to sentences with the DUC 2003 sentence segmentation tool and keep all words of length at least two.

Comparative Evaluation. We compare the Association Mixture Text Summarization method against results given in literature for state-of-the-art unsupervised summarization methods: Document Summarization Based on Data Reconstruction, linear and non-linear (DSDR-lin, DSDR-non) [5], Topic DSDR (TDSRD) [11], Two-Tiered Topic Model (TTM) and Enriched TTM (ETTM) [1]. The last two use Wordnet and topic description as additional resources. We also include two baseline methods provided with the DUC: NIST BL and CLASSY04. The latter is actually a supervised method.

4.2 Results

Association Mixture Model and Its Two Components: In terms of F-measure for ROUGE-1, Figure 1 illustrates the performance of the overall model and the independence and background corpus components as functions of the size of the background corpus $\mathcal{B}$.

The performance improves from 0.380 to 0.422 as the size of the background $\mathcal{B}$ grows from 10 to 10,000 sentences. This illustrates how a larger background corpus is a simple but effective way to provide auxiliary information to the summarization process. In our experiments, 1,000–3,000 sentences were already sufficient as a background corpus. The improvement after this was very limited.

Next, consider the performance of the two components of the model individually. The independence component does obviously not depend on the background corpus $\mathcal{B}$ and is hence represented by a horizontal line on the figure.

The background component, in turn, shows a longer period of improvement than the Association Mixture model and converges later than the 1,000–3,000 sentences range.

Overall, the Association Mixture Text Summarization method seems to successfully combine the two components into a model that clearly dominates both of them. Contrary to our expectations, there is a clear margin over the background component for large background corpus sizes, even though the relative weight of the independence component is very small there.

Comparison to Other Methods. A comparison to state-of-the-art in unsupervised summarization methods shows that the Association Mixture model is very competitive (Table 1). ROUGE-1 results are additionally shown as thin, unlabeled horizontal lines in Figure 1.

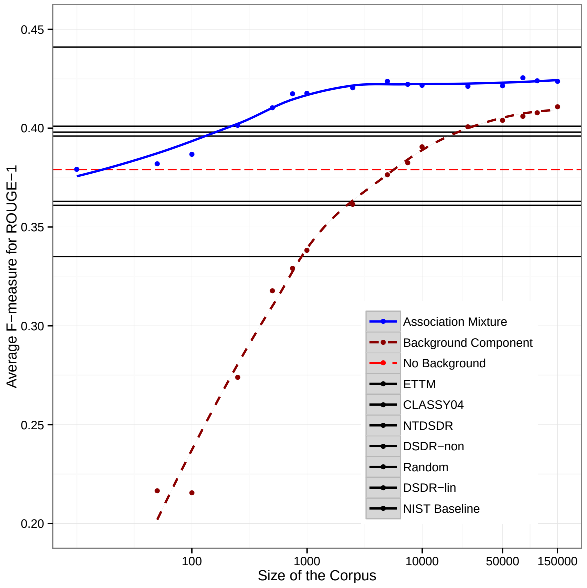

Figure 1: Performance of the methods in terms of average ROUGE-1 F-measure, as the function of the size of the background corpus $\mathcal{B}$ (smooth curves obtained by LOESS regression).

Method	Rouge-1	Rouge-2	Rouge-3	Rouge-L
NIST BL	0.335	0.065	0.019	0.311
DSDR-lin [5]	0.361	0.072	0.021	0.324
Random	0.363	0.064	0.018	0.335
DSDR-non [5]	0.396	0.074	0.020	0.353
NTDSDR [11]	0.398	0.082	-	0.362
CLASSY04	0.401	0.093	0.031	0.363
Assoc. Mix.$^+$	0.424$^+$	0.104$^+$	0.036$^+$	0.384$^+$
ETTM [1]*	0.441*	0.104*	-	-
TTM [1]*	0.447*	0.107*	-	-

Table 1: Average F measures for the DUC 2007 dataset. *Uses Wordnet and topic descriptions as additional resources. $^+$Uses background corpus as an additional resource. Paired Wilcoxon Test p-values are below 0.0004 between CLASSY04 and Assoc. Mix for all metrics.

The Association Mixture Text Summarization method outperformed all unsupervised approaches that do not rely on additional resources, and did this already with a background corpus of 300 sentences.

Among the tested methods, the Association Mixture Text Summarization method was only outperformed by the Two-Tiered Topic Models TTM and ETTM [1]. These methods use Wordnet and a topic description as additional resources, while we use a raw unprepared background corpus (with similar performance improvement with different genres and types of background corpora). It seems natural that methods using such manually crafted resources as Wordnet do better than methods using simple corpora.

5. CONCLUSIONS

In this paper we have proposed the Association Mixture Text Summarization method for creating (multi-)document summaries based on word associations. This approach has a number of characteristics: (i) it looks for relevant associations rather than words, (ii) it generalizes to multiple documents, (iii) it is unsupervised and uses simple resources, and thus it is (iv) largely language-independent.

In our experiments, the Association Mixture Text Summarization method outperformed resource-free unsupervised summarization methods and its performance was comparable to systems which use hand-crafted linguistic resources. Its performance converged when the size of the background reached approximately 1,000–3,000 sentences.

The only language-specific resource required by the method is a background corpus of some thousands of sentences, and the only required linguistic processing is the ability to split a text into sentences and its sentences into words. The simplicity of the method and its very modest requirements should make it universally applicable.

Acknowledgements.

This work has been supported by the European Commission (FET grant 611733, ConCreTe) and the Academy of Finland (Algodan Centre of Excellence).

6. REFERENCES

[1] A. Celikyilmaz and D. Hakkani-Tür. Discovery of topically coherent sentences for extractive summarization. In *ACL*, pages 491–499, 2011.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[3] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[4] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *24th international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM, 2001.

[5] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He. Document Summarization Based on Data Reconstruction. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 620–626, 2012.

[6] D. S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278, Dec. 1974.

[7] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL*, NAACL, pages 71–78. Association for Computational Linguistics, 2003.

[8] C.-Y. Lin and F. Och. Looking for a few good metrics: Rouge and its evaluation. In *NTCIR Workshop*, 2004.

[9] G. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[10] A. Nenkova and K. McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012.

[11] Z. Zhang, H. Li, et al. TopicDSDR: combining topic decomposition and data reconstruction for summarization. In *Web-Age Information Management*, pages 338–350. Springer, 2013.