Extractive Summarisation via Sentence Removal: Condensing Relevant Sentences into a Short Summary

Marco Bonzanini, Miguel Martinez-Alvarez and Thomas Roelleke
Queen Mary University of London
{marcob,miguel,thor}@eecs.qmul.ac.uk

ABSTRACT

Many on-line services allow users to describe their opinions about a product or a service through a review. In order to help other users to find out the major opinion about a given topic, without the effort to read several reviews, multi-document summarisation is required. This research proposes an approach for extractive summarisation, supporting different scoring techniques, such as cosine similarity or divergence, as a method for finding representative sentences. The main contribution of this paper is the definition of an algorithm for sentence removal, developed to maximise the score between the summary and the original document. Instead of ranking the sentences and selecting the most important ones, the algorithm iteratively removes unimportant sentences until a desired compression rate is reached. Experimental results show that variations of the sentence removal algorithm provide good performance.

Categories and Subject Descriptors

H.3.m [**Information Search and Retrieval**]: Miscellaneous

General Terms

Algorithms, Design, Experimentation

Keywords

Opinion Summarisation, Sentence Removal, Divergence

1. INTRODUCTION & MOTIVATION

The expansion of the Web has provided an increasing number of social media sources such as blogs, discussion forums and other services, where users can express their opinions about products, companies or people. Finding out what other people think has always been an important part of our decision-making process [11]. For example, customers can exploit opinion-oriented information before buying a product. In order for this information to be effective and not overwhelming, intelligent sentiment-aware tools are needed.

Automatic document summarisation is an important task which provides an effective access to information. It has been extensively explored as means to reduce the information overload [10]. The purpose of a summariser is to provide the user with the most important information from the original source, in a short form.

In the context of opinions, document summarisation can help to find out a concise way to express what the different users think about products and services. Given a set of reviews, a retrieval system could respond to information needs such as *find opinions about the sound quality of the new iPod*, but the users would still be required to read a number of sentences in order to understand what the pivot opinion is.

In order to provide a snippet representing this pivot opinion, a summariser can be joined to the aforementioned retrieval system, as a second-stage component. Figure 1 provides an overview of such a two-stage system, showing the pipeline which leads from the information need (e.g. tell me the major opinion about a topic) to the generation of a short answer to the query.

This work focuses on how an extractive summariser can provide such a short snippet from a set of redundant sentences, similarly to the scenario described above.

The main contribution of this work consists in the definition of a novel approach to summarisation, based on Sentence Removal (SR). This technique removes the less important sentences until the desired summary length is reached. With this approach, the summarisation procedure considers the importance of a candidate summary as a whole, rather than focusing on the importance of a single sentence, and tries to maximise the coverage of relevant information. Different scoring techniques can define the importance of sentences, and an experimental study which explores cosine similarity and divergence is reported.

2. BACKGROUND & RELATED WORK

Automatic document summarisation is the task of presenting a shortened version of a document, or a set of documents, containing the most important information expressed in the source. Professional human abstractors can produce high-quality summaries, but they often require domain-specific knowledge, as well as time and costs which could be unaffordable. Intelligent tools for summarisation are crucial in the process of information reduction.

2.1 Sentence Extraction

Traditional sentence extraction techniques apply different methods for determining the importance of sentences [10]. Experiments in sentence extraction based on significance scores have been reported since the 1950's. Later research has investigated the combination of different features for identifying significant sentences [4]. Word frequency has been associated with cue words presence, title and heading words presence and structural information such as sen-

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

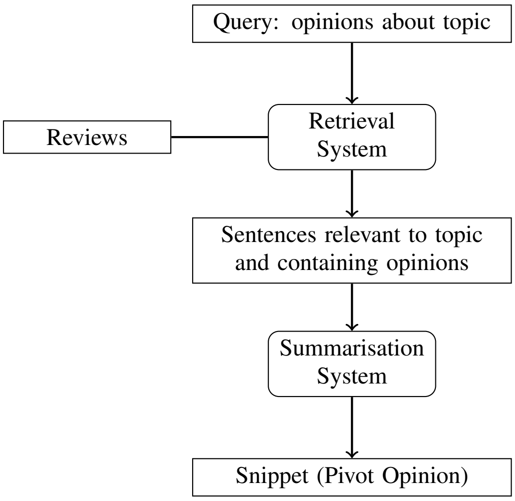
Query: opinions about topic
Reviews
Retrieval System
Sentences relevant to topic and containing opinions
Summarisation System
Snippet (Pivot Opinion)

Figure 1: Two-stage system for summarising opinions.

tence location. A weighted combination of these features is used to select the sentences to extract.

Instead of using subjective weights to combine features, a corpus can be used to train a classifier. With this approach, the summarisation task has been faced as a sentence classification task. A Naive-Bayes classifier has been used to determine whether a sentence has to be selected to generate the summary [7].

Other approaches based on word statistics include the use of TF-IDF weighting, which is commonly incorporated in most of the current summarisation systems [10], or the application of the log-likelihood ratio test for the identification of highly descriptive words, commonly referred to as topic signatures. The log-likelihood approach has been shown to be particularly effective in the task of multi-document summarisation of news [3].

2.2 Sentiment Summarisation

Previous work in summarisation of opinionated documents has been focusing on different domains of user-generated content. The idea of a single sentence extraction, to determine the polarity of the whole document, has been suggested in [1], although results on the polarity classification task have not been reported.

Dealing with short web comments, the task of extracting the top sentiment keywords has been faced exploiting Pointwise Mutual Information, and the result can be presented in a tag cloud [12].

Feature-based Sentiment Analysis extracts the different aspects of a topic, providing polarity information about it. Aggregating per-aspect sentiments easily leads to the generation feature-based summaries, but previous works in this direction have mainly focused on the aspect identification and sentiment classification tasks [6, 2].

3. EXTRACTIVE SUMMARISATION

In this section the task of extractive summarisation is formalised. A collection T has a number of topics $< t_1, ..., t_m >$. A topic t_j is composed by a number of documents $< d_1, ..., d_k >$, each of which is composed by a number of sentences. A topic is hence composed by all the sentences $< s_1, ..., s_n >$ belonging to the k documents. The case $k = 1$ is called single-document summarisation, while $k > 1$ represents multi-document summarisation. The task of extractive summarisation is to select the subset of sentences and to combine them into a summary which better represents the topic. In order to form the summary, a length limit has to be considered, based on the number of sentences or the number of words. For each sentence s belonging to a topic t, its probability distribution over terms is given by:

$$P(w|s, t, T) = \sum_{d \in t} P(w|d, T) \cdot P(d|s, T) \tag{1}$$

$P(d|s, T)$ can be obtain via the Bayesian rule:

$$P(d|s, T) = \frac{P(s|d, T) \cdot P(d)}{P(s)} \tag{2}$$

then calculating $P(s|d, T)$ as follows:

$$P(s|d, T) = \prod_{w \in s} P(w|d, T) \tag{3}$$

and considering a uniform distribution for $P(d)$ and $P(s)$.

For the estimation of $P(w|d, T)$, linear smoothing is adopted:

$$P(w|d, T) = \lambda_d \cdot P(w|d) + (1 - \lambda_d) \cdot P(w|T) \tag{4}$$

where $P(w|d)$ is the relative frequency of a term w in the document d, and $P(w|T)$ is the relative frequency of a term w in the whole collection of topics T, i.e. the background model. The parameter λ_d is defined as a Dirichlet mixture, i.e. $\frac{|d|}{|d|+\mu}$, where $|d|$ is the document length, and μ is defined as the average document length.

3.1 Sentence Selection

Traditionally, summarisation approaches build the summary by selecting the most significant sentences from the original source. In order to measure the significance of a sentence, one can employ different ranking techniques. Once the sentences are ranked, the top l sentences will form the summary.

The ranking techniques analysed in this work are based on similarity and divergence. Sentences can be ranked with the purpose of maximising their similarity with the given topic, i.e.:

$$\text{score}_{\texttt{SIM}}(s, t) := \text{sim}(P_s, P_t) \tag{5}$$

where sim can be any similarity metric, P_s is the term probability distribution for the sentence s, as defined in Equation 1, and P_t is the relative frequency distribution over the topic t. In this work, cosine similarity is employed as similarity metric. A baseline which forms the candidate summary picking the top l sentences according to Equation 5 is referred to as $\text{Greedy}_{\texttt{SIM}}$ in the experiments.

A different approach to score sentences consists in minimising a measure of dissimilarity between a sentence and the given topic. In this work, Kullback-Leibler (KL) divergence is used. KL-divergence quantifies the proximity of two probability distributions. In particular, it measures the information lost when approximating a probability distribution P with a candidate distribution Q:

$$D_{\text{KL}}(P||Q) = \sum_i P_i \cdot \log\Big(\frac{P_i}{Q_i}\Big) \tag{6}$$

Given a topic t to summarise, for each sentence s belonging to the topic, one can calculate the following score:

$$\text{score}_{\texttt{DIV}}(s, t) := -D_{\text{KL}}(P_t||P_s) \tag{7}$$

where the minus sign indicates that the lowest divergence gives the highest score. A baseline which forms the candidate summary picking the top l sentences according to Equation 7 is referred to as $\text{Greedy}_{\texttt{DIV}}$ in the experiments.

Rather than selecting sentences individually, a different possibility is to select directly the subset of sentences which maximises the chosen score. The brute-force approach consists in enumerating all the combinations of l sentences (desired output length) out of the n forming the given topic. The concatenation of the l sentences with the highest score will form the summary. This is different from selecting one sentence at a time, as the language model for the concatenation will be different from the language model for the individual sentences. The two baselines implementing the brute-force approach adopting the scores as in Equations 5 and 7 are referred to as, respectively, $\mathrm{BF}_{\texttt{SIM}}$ and $\mathrm{BF}_{\texttt{DIV}}$ in the experiments.

3.2 Sentence Removal Algorithm

This section describes the proposed approach to extractive summarisation via a Sentence Removal (SR) algorithm. Instead of selecting important sentence, the idea behind this technique is based on removing iteratively the less important ones, until the desired output size is reached. With this method, the algorithm tries to maximise the importance of the candidate summary as a whole, and does not only focus on the importance of a single sentence. In other words, the purpose is to condense the information in the original source while trying to ensuring its coverage in the summary at the same time. Algorithm 1 shows the procedure to obtain the candidate summary. The procedure starts with the candidate summary t' containing all the original set of sentences, and then iterates until the summary reaches the desired length l. During each iteration, the procedure removes one sentence such that the score between the candidate summary and the original set of sentences is maximised. The score in line 6 can be computed using again Equations 5 or 7 In the results section, the systems implementing the SR algorithm as in Algorithm 1 are denoted with $\mathrm{SR}_{\texttt{SIM}}$ and $\mathrm{SR}_{\texttt{DIV}}$ depending on the scoring function.

Algorithm 1 Sentence Removal algorithm.

```
Input: $t$ {topic to summarise}
Input: $l$ {output size in n. of sentences}
1: $t' \leftarrow t$
2: while $|t'| > l$ do
3:     for all $s_i$ in $t'$ do
4:         $t'_i \leftarrow t' \setminus s_i$
5:     end for
6:     $t' \leftarrow \arg\max_{t'_i}(\mathrm{score}(t, t'_i))$
7: end while
8: return $t'$
```

A different version of the sentence removal algorithm can be obtained with a variation in the way the candidate summary, at each iteration step, is selected. Rather than computing the score between the candidate summary and the original set of sentences, one can compute the score between the candidate summary t'_i, and the candidate summary at the previous iteration t'. In this case, line 6 of the procedure has to be replaced with:

$$t' \leftarrow \arg\max_{t'_i}(\mathrm{score}(t', t'_i))$$

The systems implementing this variation of the algorithm are labelled as $\mathrm{SR}'_{\texttt{SIM}}$ and $\mathrm{SR}'_{\texttt{DIV}}$. This version of the algorithm could also be exploited to shift the topic of the summary towards a particular sub-topic, e.g. via a topic query.

4. EXPERIMENTAL STUDY

An experimental study has been performed using a dataset suitable for multi-document summarisation. The dataset provides a number of documents categorised into topics, as well as golden standard summaries for each topic.

Recall			
	ROUGE-1	ROUGE-2	ROUGE-SU4
MEAD	**49.32** †	**10.58**	**23.16** †
Greedy$_{\texttt{SIM}}$	33.80	06.43	12.12
BF$_{\texttt{SIM}}$	22.32	05.75	05.43
SR$_{\texttt{SIM}}$	*37.46*	*09.29*	*13.80*
SR$'_{\texttt{SIM}}$	15.78	02.64	03.03
Greedy$_{\texttt{DIV}}$	18.84	03.99	03.91
BF$_{\texttt{DIV}}$	20.46	05.54	04.97
SR$_{\texttt{DIV}}$	*46.05*	*08.67*	*20.10*
SR$'_{\texttt{DIV}}$	15.60	01.44	02.96
Precision			
	ROUGE-1	ROUGE-2	ROUGE-SU4
MEAD	09.16	01.84	01.02
Greedy$_{\texttt{SIM}}$	15.21	02.78	03.00
BF$_{\texttt{SIM}}$	*29.39*	*07.78*	*10.27*
SR$_{\texttt{SIM}}$	19.87	05.18	05.44
SR$'_{\texttt{SIM}}$	25.36	04.68	08.72
Greedy$_{\texttt{DIV}}$	30.42	06.64	11.22
BF$_{\texttt{DIV}}$	**30.70**	**08.36**	**12.10**
SR$_{\texttt{DIV}}$	09.64	01.77	01.10
SR$'_{\texttt{DIV}}$	12.70	01.20	02.23
F_1**-score**			
	ROUGE-1	ROUGE-2	ROUGE-SU4
MEAD	15.15	03.08	01.89
Greedy$_{\texttt{SIM}}$	19.80	03.66	04.19
BF$_{\texttt{SIM}}$	**24.67**	*06.39*	*06.42*
SR$_{\texttt{SIM}}$	24.38	06.23	06.31
SR$'_{\texttt{SIM}}$	19.01	03.28	04.16
Greedy$_{\texttt{DIV}}$	22.84	04.88	05.47
BF$_{\texttt{DIV}}$	*24.03*	**06.50**	**06.59**
SR$_{\texttt{DIV}}$	15.64	02.88	02.03
SR$'_{\texttt{DIV}}$	13.33	01.25	02.11
F_2**-score**			
	ROUGE-1	ROUGE-2	ROUGE-SU4
MEAD	26.27	05.43	04.34
Greedy$_{\texttt{SIM}}$	25.19	04.71	05.94
BF$_{\texttt{SIM}}$	23.05	05.95	05.66
SR$_{\texttt{SIM}}$	**29.92** †	**07.54** †	**08.28** †
SR$'_{\texttt{SIM}}$	16.88	02.86	03.37
Greedy$_{\texttt{DIV}}$	20.20	04.29	04.39
BF$_{\texttt{DIV}}$	21.68	*05.86*	*05.48*
SR$_{\texttt{DIV}}$	*25.39*	04.70	04.16
SR$'_{\texttt{DIV}}$	14.38	01.34	02.38

Table 1: ROUGE scores on the Opinosis dataset. The best overall results are shown in bold. The best results within the same scoring function are shown in *italic*. Best results labelled with a † show that the second-best results are outside their 95% confidence interval.

The Opinosis dataset [5] is a collection of opinion-oriented data, divided into 51 different topics. Each topic represents an aspect of a product or service, for example *Battery Life of the Amazon Kindle*, or *Food quality of the Holiday Inn London*. Each topic includes a number of sentences (min. 50, max. 575, avg. 139), taken from different reviews from popular review web sites. For each topic, 4 or 5 golden standard (human-written) summaries are provided. The data are not labelled according to their sentiment, and different opinions can be expressed. The golden standard summaries hence present the pivot opinion for each topic, in a concise way (approx. 2 sentences each). For this reason, the maximum length of the system generated summaries is fixed to two sentences. Another characteristic of the data is to be highly redundant, e.g. the topic itself is often repeated in different sentences.

The ROUGE framework [9] is used to provide a quantitative assessment between the candidate summaries and the golden standards. Specifically, the results for ROUGE-1, ROUGE-2 and ROUGE-SU4 are reported. In order to achieve better correlation with human judgement, multiple golden standard summaries are used [8]. Given the brevity of the summaries, capturing all the relevant information is particularly challenging. Recall is hence particularly important, i.e. it is desirable to show *all* the relevant opinions in the summaries, yet maintaining their succinct nature. For this reason, the results for F_2-scores, which emphasise the importance of recall over precision, are also reported. On top of the baselines described in Section 3.1, this study also reports the results for MEAD [13], a state-of-the-art extractive summariser based on cluster centroids.

4.1 Results

Table 1 reports the results of the experiments on the Opinosis data set. The MEAD baseline is extremely competitive with respect to recall, but shows a drop of performance on the precision side. Overall, the results show that there is not an individual approach which clearly outperforms all the others in every metric. In general, the variation of the sentence removal algorithm, SR', is outperformed by the original definition of SR, for both the cosine and the divergence-based settings. The SR algorithm consistently achieves the best recall results within the same scoring function groups when compared to greedy or brute-force approaches. Observing the cosine-based systems, the brute-force approach shows the best F_1-scores, but its results are not substantially better than the SR algorithm. They both outperform the greedy baseline and MEAD. On the F_2-scores side, SR is substantially better than any other system, including the divergence-based ones and MEAD, with the second-best results being outside a 95% confidence interval. On the divergence-based side, the brute-force approach achieves the best F_1-scores due to good performance in precision. Its F_2-scores are slightly better than the SR ones for ROUGE-2 and ROUGE-SU4, while SR achieves the best ROUGE-1 score for the divergence-based systems.

5. DISCUSSION & CONCLUSIONS

This paper discussed the use of a summariser as a component of a two-stage system, which shows opinion-oriented summaries as a result of an opinion-oriented query. Such a summariser is meant to take a set of relevant sentences and condense them into a short summary. The main contribution is the introduction of a novel algorithm for extractive summarisation based on Sentence Removal (SR). The key idea is to remove unimportant sentences iteratively, until a desired output length is reached, rather then selecting directly the important ones. In order to define the importance of a sentence, a scoring function can be applied. In this work, scores based on cosine similarity and on divergence have been investigated. An experimental study has compared the performance of the SR algorithm against MEAD, a state-of-the-art summariser, and against two baselines, namely greedy and brute-force approaches which adopt the same scoring function as the SR algorithm. Results do not conclusively assert which scoring function is overall better (cosine vs. divergence), and provide promising indication of the effectiveness of the SR algorithm. In particular, the SR algorithm is comparable to the best baseline in terms of F_1-scores, and it is overall the best approach in terms of F_2-scores. Given the nature of summarisation, and in particular, opinion summarisation, future work includes the investigation of an alternative representation for sentences, documents and topics. In particular the hypothesis is that a representation based on subspaces can help to capture opinions better than a single-dimensional one, because it directly enables to treat different groups of terms, e.g. opinion-bearing words, in different ways.

6. REFERENCES

[1] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. Exploring sentiment summarization. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applicationsi*, 2004.

[2] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G.A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, 2008.

[3] J.M. Conroy, J.D. Schlesinger, and D.P. O'Leary. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL2006*, pages 152–159, 2006.

[4] H.P. Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285, 1969.

[5] K. Ganesan, C.X. Zhai, and J. Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd COLING*, pages 340–348, 2010.

[6] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence*, pages 755–760. AAAI, 2004.

[7] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th ACM SIGIR*, pages 68–73, 1995.

[8] C.Y. Lin. Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough. In *Proceedings of the 4th NTCIR Workshop*, pages 1–10, 2004.

[9] C.Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81, 2004.

[10] A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 2011.

[11] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[12] M. Potthast and S. Becker. Opinion Summarization of Web Comments. In *Proceedings of the 32nd ECIR*, pages 668–669, 2010.

[13] Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. MEAD - a platform for multidocument multilingual text summarization. In *LREC*, 2004.