Detecting Novelty in the context of Progressive Summarization

Praveen Bysani
Language Technologies Research Center
IIIT Hyderabad
lvsnpraveen@research.iiit.ac.in

Abstract

A Progressive summary helps a user to monitor changes in evolving news topics over a period of time. Detecting novel information is the essential part of progressive summarization that differentiates it from normal multi document summarization. In this work, we explore the possibility of detecting novelty at various stages of summarization. New scoring features, Re-ranking criterions and filtering strategies are proposed to identify "relevant novel" information. We compare these techniques using an automated evaluation framework ROUGE, and determine the best. Overall, our summarizer is able to perform on par with existing prime methods in progressive summarization.

1 Introduction

Summarization is the process of condensing text to its most essential facts. Summarization is challenging for its associated cognitive task and interesting because of its practical usage. It has been successfully applied for text content such as news articles [1], scientific papers (Teufel and Moens, 2002) that follow a discourse structure. Update summarization is an emerging area with in summarization, acquiring significant research focus during recent times. The task was introduced at DUC 2007[2] and continued during TAC 2008, 2009[3]. We refer to update summariztion as "Progressive Summarization" in rest of this paper, as summaries are produced periodically in a progressive manner and the latter title is more apt to the task. Progressive summaries contain information which is both relevant and novel, since they are produced under the assumption that user has already read some previous documents/articles on the topic. Such summaries are extremely useful in tracking news stories, tracing new product reviews etc.

Unlike dynamic summarization (Jatowt, 2004) where a single summary transforms periodically, reflecting changes in source text, Progressive summarizer produce multiple summaries at specific time intervals updating user knowledge. Temporal Summarization (Allan et al., 2001) generate summaries, similar to progressive summaries by ranking sentences as combination of *relevant and new* scores. In this work, summaries are produced not just by reforming ranking scheme but also altering scoring and extraction stages of summarization.

Progressive summarization requires differentiating *Relevant and Novel* Vs *Non-Relevant and Novel* Vs *Relevant and Redundant* information. Such discrimination is feasible only with efficient Novelty detection techniques. We define Novelty detection as identifying relevant sentences containing new information. This task shares similarity with TREC Novelty Track [4], that is designed to investigate systems abilities to locate sentences containing relevant and/or new information given the topic and a set of relevant documents ordered by date. A progressive summarizer needs to identify, score and then finally rank "relevant novel" sentences to produce a summary.

[1] http://newsblaster.cs.columbia.edu/

[2] http://duc.nist.gov/duc2007/tasks.html

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

[4] http://trec.nist.gov/data/novelty.html

Proceedings of the NAACL HLT 2010 Student Research Workshop, pages 13–18,
Los Angeles, California, June 2010.

Previous approaches to Novelty detection at TREC (Soboroff, 2004) include cosine filtering (Abdul-Jaleel et al., 2004), where a sentence having maximum cosine similarity value with previous set of sentences, lower than a preset threshold is considered novel. Alternatively, (Schiffman and McKeown, 2004) considered previously unseen words as an evidence of Novelty. (Eichmannac et al., 2004) expanded all noun phrases in a sentence using wordnet and used corresponding sysnsets for novelty comparisions.

Our work targets exploring the effect of detecting novelty at different stages of summarization on the quality of progressive summaries. Unlike most of the previous work (Li et al., 2009) (Zhang et al., 2009) in progressive summarization, we employ multiple novelty detection techniques at different stages and analyze them all to find the best.

2 Document Summarization

The Focus of this paper is only on extractive summarization, henceforth term summarization/summarizer implies sentence extractive multi document summarization. Our Summarizer has 4 major stages as shown in Figure 1,

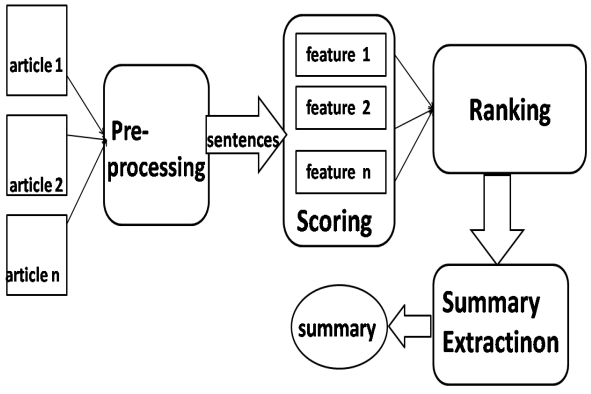

Figure 1: Stages in a Multi Document Summarizer

Every news article/document is cleaned from news heads, HTML tags and split into sentences during *Pre-processing* stage. At *scoring*, several sentence scoring features assign scores for each sentence, reflecting its topic relevance. Feature scores are combined to get a final rank for the sentence in *ranking* stage. Rank of a sentence is predicted from regression model built on feature vectors of sentences in the training data using support vector machine as explained in (Schilder and Kondadandi, 2008). Finally during *summary extraction*, a subset of ranked sentences are selected to produce summary after a redundancy check to filter duplicate sentences.

2.1 Normal Summarizers

Two normal summarizers (*DocSumm, TacBaseline*) are developed in a similar fashion described in Figure 1.

DocSumm produce summaries with two scoring features, Document Frequency Score (DF) (Schilder and Kondadandi, 2008) and Sentence Position (SP). DocSumm serves as a baseline to depict the effect of novelty detection techniques described in Section 3 on normal summarizers. *Document frequency (DF)*, of a word (w) in the document set ($docs$) is defined as ratio of number of documents in which it occured to the total number of documents. Normalized DF score of all content words in a sentence is considered its feature score.

$$DF_{docs}(w) = \frac{\{|d| : w \in d\}}{|docs|}$$

Sentence Position (SP) assigns positional index (n) of a sentence (s_n) in the document (d) it occurs as its feature score. Training model will learn the optimum sentence position for the dataset.

$$SP(s_{nd}) = n$$

TacBaseline is a conventional baseline at TAC, that creates a n word length summary from first n words of the most recent article. It provides a lower bound on what can be achieved with automatic multi document summarizers.

3 Novelty Detection

Progressive summaries are generated at regular time intervals to update user knowledge on a particular news topic. Imagine a set of articles published on a evolving news topic over time period T, with t_d being publishing timestamp of article d. All the articles published from time 0 to time t are assumed to

have been read previously, hence prior knowledge, *pdocs*. Articles published in the interval t to T that contain new information are considered *ndocs*.

$$ndocs = \{d : t_d > t\}$$
$$pdocs = \{d : t_d <= t\}$$

Progressive summarization needs a novelty detection technique to identify sentences that contain relevant new information. The task of detecting novelty can be carried out at 3 stages of summarization shown in Figure 1.

3.1 At Scoring

New Sentence scoring features are devised to capture sentence novelty along with its relevance. Two features Novelty Factor (NF) (Varma et al., 2009), and New Words (NW) are used at scoring level.

Novelty Factor (NF)
NF measures both topic relevancy of a sentence and its novelty given prior knowledge of the user through $pdocs$. NF score for a word w is calculated as,

$$NF(w) = \frac{|nd_t|}{|pd_t| + |ndocs|}$$

$nd_t = \{d : w \in d \wedge d \in ndocs\}$
$pd_t = \{d : w \in d \wedge d \in pdocs\}$
$|nd_t|$ captures the relevancy of w, and $|pd_t|$ elevates the novelty by penalizing words occurring frequently in $pdocs$. Score of a sentence is the average NF value of its content words.

New Words (NW)
Unlike NF, NW captures only novelty of a sentence. Novelty of a sentence is assessed by the amount of *new* words it contains. Words that never occurred before in $pdocs$ are considered *new*. Normalized term frequency of a word (w) is used in calculating feature score of sentence. Score of a sentence(s) is given by,

$$\begin{aligned} Score(s) &= \frac{\sum_{w \in s} NW(w)}{|s|} \\ NW(w) &= 0 \quad if\ w \in pdocs \\ &= n/N \quad else \end{aligned}$$

n is frequency of w in $ndocs$
N is total term frequency of $ndocs$

3.2 At Ranking

Ranked sentence set is re-ordered using Maximal Marginal relevance (Carbonell and Goldstein, 1998) criterion, such that prior knowledge is neglected and sentences with new information are promoted in the ranked list. Final rank ("Rank") of a sentence is computed as,

$$\begin{aligned} Rank &= relweight * rank - \\ &\quad (1 - relweight) * redundancy\_score \end{aligned}$$

Where "rank" is the original sentence rank predicted by regression model as described in section 2, and "redundancy_score" is an estimate for the amount of prior information a sentence contains. Parameter "relweight" adjusts relevancy and novelty of a sentence. Two similarity measures *ITSim, CoSim* are used for calculating redundancy_score.

Information Theoretic Similarity (ITSim)
According to information theory, Entropy quantifies the amount of information carried with a message. Extending this analogy to text content, Entropy $I(w)$ of a word w is calculated as,

$$I(w) = -p(w) * log(p(w))$$
$$p(w) = n/N$$

Motivated by the information theoretic definition of similarity by (Lin, 1998), we define similarity between two sentences $s1$ and $s2$ as,

$$ITSim(s1, s2) = \frac{2 * \sum_{w \in s1 \wedge s2} I(w)}{\sum_{w \in s1} I(w) + \sum_{w \in s2} I(w)}$$

Numerator is proportional to the commonality between $s1$ and $s2$ and denominator reflects differences between them.

Cosine Similarity (CoSim)
Cosine similarity is a popular technique in TREC Novelty track to compute sentence similarity. Sentences are viewed as tf-idf vectors (Salton and Buckley, 1987) of words they contain in a *n-dimension* space. Similarity between two sentences is measured as,

$$CoSim(s1, s2) = cos(\Theta) = \frac{s1.s2}{|s1||s2|}$$

Average similarity value of a sentence with all sentences in $pdocs$ is considered as its redundancy score.

3.3 At summary extraction

Novelty Pool (NP)
Sentences that possibly contain prior information are filtered out from summary by creating Novelty Pool (NP), a pool of sentences containing one or more $novelwords$. Two sets of "dominant" words are generated one for each $pdocs$ and $ndocs$.

$$dom_{ndocs} = \{w : DF_{ndocs}(w) > threshold\}$$

$$dom_{pdocs} = \{w : DF_{pdocs}(w) > threshold\}$$

A word is considered dominant if it appears in more than a predefined "threshold " of articles, thus measuring its topic relevance. Difference of the two dom sets gives us a list of $novelwords$ that are both relevant and new.

$$novelwords = dom_{ndocs} - dom_{pdocs}$$

4 Experiments and Results

We conducted all the experiments on TAC 2009 Update Summarization dataset. It consists of 48 topics, each having 20 documents divided into two clusters "A" and "B" based on their chronological coverage of topic. It serves as an ideal setting for evaluating our progressive summaries. Summary for cluster A (*pdocs*) is a normal multi document summary where as summary for cluster B (*ndocs*) is a Progressive summary, both of length 100 words. Each topic has associated 4 model summaries written by human assessors. TAC 2008 Update summarization data that follow similar structure is used to build training model for support vectors as mentioned in Section 2. Thresholds for dom_{ndocs}, dom_{pdocs} are set to 0.6, 0.3 respectively and *relweight* to 0.8 for optimal results.

Summaries are evaluated using ROUGE (Lin, 2004), a recall oriented metric that automatically assess machine generated summaries based on their overlap with models. ROUGE-2 and ROUGE-SU4 are standard measures for automated summary evaluation. In Table 1 ROUGE scores of baseline systems(Section 2.1) are presented.
Five progressive runs are generated, each having a novelty detection scheme at either scoring, ranking or summary extraction stages. ROUGE scores of these runs are presented in Table 2.

	ROUGE-2	**ROUGE-SU4**
DocSumm	0.09346	0.13233
TacBaseline	0.05865	0.09333

Table 1: Average ROUGE-2, ROUGE-SU4 recall scores of $baselines$ for TAC 2009, cluster B

NF+DocSumm : Sentence scoring is done with an additional feature NF, along with default features of DocSumm
NW+DocSumm : An additional feature NW is used to score sentences for DocSumm
ITSim+DocSumm : ITSim is used for computing similarity between a sentence in *ndocs* and set of all sentences in *pdocs*. Maximum similarity value is considered as redundancy_score. Re-ordered ranked list is used for summary extraction
Cosim+DocSumm : CoSim is used as a similarity measure instead of ITSim
NP+DocSumm : Only members of NP are considered while extracting DocSumm summaries

Results of top systems at TAC 2009, *ICSI* (Gillick et al., 2009) and *THUSUM* (Long et al., 2009) are also provided for comparison.

	ROUGE-2	**ROUGE-SU4**
ICSI	0.10417	0.13959
NF+DocSumm	0.10273	0.13922
NW+DocSumm	0.09645	0.13955
NP+DocSumm	0.09873	0.13977
THUSUM	0.09608	0.13499
ITSim+DocSumm	0.09461	0.13306
Cosim+DocSumm	0.08338	0.12607

Table 2: Average ROUGE-2, ROUGE-SU4 recall scores for TAC 2009, cluster B

Next level of experiments are carried out on combination of these techniques. Each run is produced by combining two or more of the above(Section 3) described techniques in conjunction with $DocSumm$. Results of these runs are presented in table 3
NF+NW : Both NF and NW are used for sentence scoring along with default features of DocSumm
NF+NW+ITSim : Sentences scored in NF+NW are re-ranked by their ITSim score
NF+NW+NP : Only members of NP are selected while extracting NF+NW summaries

NF+NW+ITSim+NP : Sentences are selected from NP during extraction of NF+NW+ITSim summaries

	ROUGE-2	ROUGE-SU4
NF+NW	0.09807	0.14058
NF+NW+ITSim	0.09704	0.13978
NF+NW+NP	0.09875	0.14010
{NP+NW+ ITSim+NP}	0.09664	0.13812

Table 3: Average ROUGE-2, ROUGE-SU4 recall scores for TAC 2009, cluster B

5 Conclusion and Discussion

Experimental results prove that proposed Novelty Detection techniques, particularly at scoring stage are very effective in the context of progressive summarization. Both NF, a language modeling technique and NW, a heuristic based feature are able to capture relevant novelty successfully. An approximate 6% increase in ROUGE-2 and 3% increase in ROUGE-SU4 scores over DocSumm support our argument. Scores of NF+DocSumm and NW+DocSumm are comparable with existing best approaches. Since CoSim is a word overlap measure, and novel information is often embedded within a sentence containing formerly known information, quality of progressive summaries declined. ITSim performs better than Cosim because it considers entropy of a word in similarity computations, which is a better estimate of information. There is a need for improved similarity measures that can capture semantic relatedness between sentences. Novelty pool (NP) is a simple filtering technique, that improved quality of progressive summaries by discarding probable redundant sentences into summary. From the results in Table 2, it can be hypothesized that Novelty is best captured at sentence scoring stage of summarization, rather than at ranking or summary extraction.

A slight improvement of ROUGE scores is observed in table 3, when novelty detection techniques at scoring, ranking and extracting stages are combined together. As Novel sentences are already scored high through NF and NW, the effect of Re-Ranking and Filtering is not significant in the combination.

The major contribution of this work is to identify the possibility of novelty detection at different stages of summarization. Two new sentence scoring features (NF and NW), a filtering strategy (NP), a sentence similarity measure (ITSim) are introduced to capture relevant novelty. Although proposed approaches are simple, we hope that this novel treatment could inspire new methodologies in progressive summarization. Nevertheless, the problem of progressive summarization is far from being solved given the complexity involved in novelty detection.

Acknowledgements

I would like to thank Dr. Vasudeva Varma at IIIT Hyderabad, for his support and guidance throughout this work. I also thank Rahul Katragadda at Yahoo Research and other anonymous reviewers, for their valuable suggestions and comments.

References

Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, and Xiaoyan Li. 2004. Umass at trec 2004: Novelty and hard.

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of news topics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA. ACM.

David Eichmannac, Yi Zhangb, Shannon Bradshawbc, Xin Ying Qiub, Padmini Srinivasanabc, and Aditya Kumar. 2004. Novelty, question answering and genomics: The university of iowa response.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The icsi/utd summarization system at tac 2009.

Adam Jatowt. 2004. Web page summarization using dynamic content. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers and posters*, pages 344–345, New York, NY, USA. ACM.

Sujian Li, Wei Wang, and Yongwei Zhang. 2009. Tac 2009 update summarization with unsupervised methods.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learn-*

ing, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Chong Long, Minlie Huang, and Xiaoyan Zhu. 2009. Tsinghua university at tac 2009: Summarizing multi-documents by information distance.

Gerard Salton and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA.

Barry Schiffman and Kathleen R. McKeown. 2004. Columbia university in the novelty track at trec 2004.

Frank Schilder and Ravikumar Kondadandi. 2008. Fastsum: fast and accurate query-based multi-document summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*. Human Language Technology Conference.

Ian Soboroff. 2004. Overview of the trec 2004 novelty track. National Institute of Standards and Technology,Gaithersburg, MD 20899.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445.

Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharat, Santosh GSK, Karuna Kumar, Sudheer Kovelamudi, Kiran Kumar N, and Nitin Maganti. 2009. iiit hyderabad at tac 2009. Technical report, Gaithersburg, Maryland USA.

Jin Zhang, Pan Du, Hongbo Xu, and Xueqi Cheng. 2009. Ictgrasper at tac2009: Temporal preferred update summarization.