Headline Summarization at ISI

Liang Zhou and Eduard Hovy
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{liangz, hovy}@isi.edu

Abstract

Headlines are useful for users who only need information on the main topics of a story. We present a headline summarization system that is built at ISI for this purpose and is a top performer for DUC2003's task 1, generating very short summaries (10 words or less).

1 Introduction

Most previous work on summarization focused on generating summary by extraction, which is finding a subset of the document that is indicative of its contents (Kupiec et al., 1995). Still the size of extraction summaries is quite large when the sole purpose is to decide whether a document is of interest. In such scenarios, central idea(s) presented in a list of phrases (10 words) should suffice.

We introduce a headline generation system that selects headline words throughout the entire text, and then composes them by finding phrase clusters locally in the beginning of the text. After going through post-processing phase, the phrase clusters will be the resulting headline. This system participated in task 1 of DUC2003 and was one of the top performers reflected in the computation of Length Adjusted Coverage scores.

In this paper, we first describe the corpus used for training, the study on bag-of-words models on headlines in Section 3, candidate word selection, overlapping of the top-scoring words and words/phrase clustering in Section 4, post-processing on system generated headlines in Section 5, and evaluation results from DUC2003 in Section 6.

2 Training Corpus

We downloaded the Yahoo Full Coverage Collection (YFCC) from http://fullcoverage.yahoo.com during December 2001. The full coverage texts were downloaded based on a snapshot of the links contained in Yahoo Full Coverage at that time. A spider crawled the top eight categories: U.S., World, Business, Technology, Science, Health, Entertainment, and Sports. All news links in each category were saved in an index page that contained the headline and its full text URL. A page fetcher then downloaded all the pages listed in the snapshot index file.

Under the eight categories, there are 216590 news articles with their respective headlines. Since headlines generated by our system are all words/phrases extracted from the body of the articles, we reduced YFCC to only 60933 articles, each of which contains all of its headline words (stop words excluded).

3 Model Selection

3. 1 Bag-of-words Models

To imitate the art of generating headlines, we studied several bag-of-words models. Based on the statistics on various model combinations, we selected best for headline words selection.

1) Sentence Position Model*:* Sentence position information has long proven useful in identifying

topics of texts (Lin and Hovy, 1997). We believe this theory also applies to the process of selecting headline words based on their sentence position. Given a sentence with its position in text, how likely it would contain a headline word's first appearance:

$$Count\_Pos_i = \Sigma_{k=1}^{M} \Sigma_{j=1}^{N} P(H_k \mid W_j)$$

$$P(H \mid Pos_i) = Count\_Pos_i / \Sigma_{i=1}^{Q} Count\_Pos_Q$$

For each sentence number i over all M texts in the corpus, and for each word in the headline which contains N words over all M headlines, *Count_Pos* records the total number of times that headline words' first appearance is at i. $P(H_k \mid W_j)$ is a binary feature, meaning it can only be 1 or 0. This is computed for all sentence positions up to position Q.

2) Headline Word Position Model: For each headline word, which sentence position in the text it would most likely make its first appearance:

$$P(Pos_i|W_h)=Count(Pos_i,W_h)/\Sigma_{i=1}^{Q} count(Pos_Q,W_h)$$

3) Text Model: This model captures the correlation between the words in the text and the words in the headline (Lin and Hauptmann, 2001):

$$P(H_w|T_w) = \Sigma_{j=1}^{M}(doc\_tf(w,j) \times title\_tf(w,j)) / \Sigma_{j=1}^{M} doc\_tf(w, j)$$

doc_tf(w,j) denotes the term frequency of word w in the j^{th} document of all M documents in corpus. *title_tf(w,j)* is the term frequency of word w in the j^{th} title. H_w and T_w are the same word.

4) Unigram Headline Model: unigram probabilities of the headline words on headlines in the corpus alone.

5) Bigram Headline Model: bigram probabilities of the headline words on headlines.

3.2 Statistics on Model Combinations

After training on above five bag-of-words models individually, we want to see which model or model combination is best suited for headline word selection. The blind test data is 108 texts with headlines from the DUC01 testing corpus. The gold standard used in the evaluation is the provided headlines. All possible model combinations are evaluated based on word overlaps between top-scoring n words and the gold standards. Models are numbered as in Section 3.1. Table 1 shows the effectiveness of each model/model combination on top 10, 20, 30, 40, and 50 scoring words.

total headline words: 808 from 108 test files

	10-word	20-word	30-word	40-word	50-word
1 2 3 4 5	79	118	147	189	216
2 3 4 5	74	110	145	178	206
1 3 4 5	74	116	146	176	208
1 2 4 5	63	99	144	176	202
1 2 3 5	87	122	155	187	223
1 2 3 4	96	149	187	214	230
3 4 5	61	103	134	170	199
2 4 5	54	94	137	168	192
2 3 5	82	117	148	183	212
2 3 4	67	119	167	192	217
1 4 5	55	101	126	149	193
1 3 5	84	113	144	181	216
1 3 4	97	144	186	212	234
1 2 5	70	102	146	179	208
1 4 5	55	101	126	149	193
1 2 3	131	181	205	230	250
4 5	46	84	117	140	182
3 5	72	107	134	166	204
3 4	58	103	136	165	196
2 5	62	96	135	172	204
2 4	38	80	114	144	179
2 3	100	150	187	215	235
1 5	72	98	130	158	203
1 4	69	111	144	169	193
1 3	**154**	**204**	**244**	**271**	**292**
1 2	74	138	174	199	232
5	58	84	114	140	171
4	35	60	87	111	136
3	86	137	169	208	227
2	45	94	135	163	197
1	**113**	**234**	**275**	**298**	**310**

Table 1. Statistics on model combinations

Clearly, sentence position (model 1) plays the most important role in selecting headline words. Selecting the top 50 words solely based on position information practically means sentences in the beginning of a text are most informative, which is exactly the theory reflected from (Lin and Hovy, 1997). However, when we restrict the length of the generated headlines to 10 words, text model (model 3) adds its advantage on top of the position model. Since our experiment is to generate headlines with only 10 words or fewer, the combination of sentence position and text model is used for the headline word selection phase:

$$P(H|W_i) = P(H|Pos_i) \times P(Hw_i|Tw_i)$$

4 Headline Formulation

4.1 Word Selection

Section 3 explains how we can select headline-worthy words. We now have to compose them into readable headlines. As illustrated in Table1, news story headlines mostly use words from the

beginning of the text, also stated in (Zajic et al., 2002). We therefore search for n-gram phrases comprising these words in the first part of the story. Using the model combination selected in Section 3, 10 top-scoring words over the whole story are selected and highlighted in the first 50 words of the text. Figure 1 shows an example of the placement of the top-scored words in the beginning of the original text.

allegations of police racism and brutality have shaken this city that for decades has prided itself on a progressive attitude toward civil rights and a reputation for racial harmony. The death of two blacks at a drug raid that went awry, followed 10 days later by a scuffle between police and

Figure 1. Top-scoring words in text.

4.2 Clustering Algorithm

Unfortunately these words taken together do not satisfy the requirement of grammaticality. Our idea is to have the ability of pulling out the largest window of words to form the headline. As described in Section 4.1, top-scoring words appeared in the first 50 words of the text are queued for clustering. To help achieving grammaticality, we form surrounding bigrams for each word waiting in queue, shown in Figure 2.

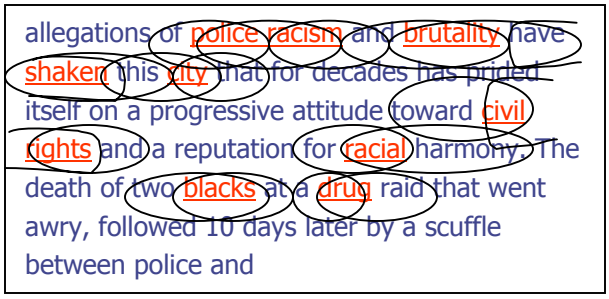

Figure 2. Surrounding bigrams for top-scoring words.

After drawing bigrams centered with top-scoring words, one can clearly see clusters of words forming. We think that the most indicative phrase for the entire text is the bigram window that has the largest number of overlapping bigrams. Since we are allowed to assert more than one topic for each headline described by the task definition, and if the generated headline length is under the limit (10 words) when one bigram window is selected, then the window with the next largest number of overlapping bigrams is appended to the headline, provided with this addition the total number of words still adheres to the length restriction. If the proposed next window is too short (less than 5 words) and the end of phrase mark (punctuation) is far away, this window will not be considered and the next largest bigram window will be examined.

5 Post Processing

The headlines generated from the above steps often contain dangling verbs, particles, etc., at the beginning and/or at the end. In order to improve the readability and the quality of the headlines, a part-of-speech tagger is run on all input texts. Using a set of hand-written rules, dangling words and words in the stop list are removed. Figure 3 shows the final generated headline on the example.

- allegations of police racism and brutality have shaken this city that for decades has prided itself on a progressive attitude toward civil rights and a reputation for racial harmony. The death of two blacks at a drug raid that went awry, followed 10 days later by a scuffle between police and
- Generated headline: police racism and brutality have shaken this city (8 words)
- Headline provided with the text (gold standard): city image tarnished by allegations of police racism (also 8 words)

Figure 3. Resulting headline.

More examples are shown in Figure 4 and Figure 5.

6 DUC Evaluation

Our system (system 13) participated in DUC2003 for Task 1 (generating very short summaries). Since it has the decision-making

ability of not appending more topics when the number of words in the generated headline gets

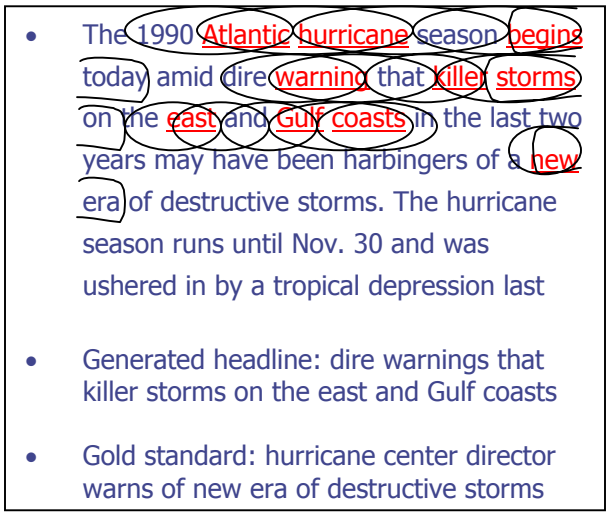

Figure 4. A good example.

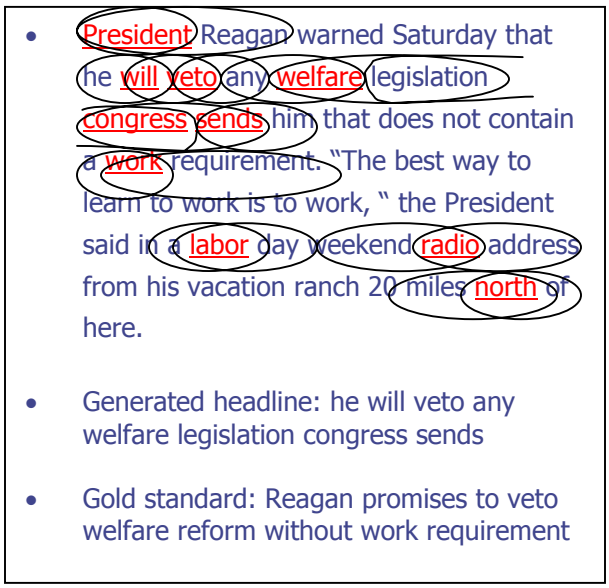

Figure 5. An example with the lack of linking *he* and *Reagan*.

dangerously close to the length restriction, the system was one of the top performers ranked by Length Adjusted Coverage measure. Table 2 shows the result of evaluation done by NIST using SEE. AVGLAC column is the evaluation results on length-adjusted coverage.

SYSID	AVGC	CNT	MEDC	AVGLAC	MEDLAC	UMPUS	CNTNZQ	AVGNZQ	AvgPeerSize
H	0.545	1872	0.545	0.367	0.367	0.034	0	0	11.018
1	0.48	624	0.48	0.375	0.375	0.058	0	0	8.651
17	0.402	624	0.402	0.276	0.276	0.123	0	0	12.593
13	0.247	624	0.247	0.264	0.264	0.095	0	0	7.03
26	0.378	624	0.378	0.258	0.258	0.037	0	0	9.833
22	0.306	624	0.306	0.252	0.252	0.099	0	0	10.663
9	0.275	624	0.275	0.248	0.248	0.107	0	0	8.274
18	0.358	624	0.358	0.247	0.247	0.072	0	0	13.96
21	0.318	624	0.318	0.212	0.212	0.091	0	0	10.508
24	0.235	564	0.235	0.232	0.232	0.096	0	0	8.388
7	0.3	624	0.3	0.207	0.207	0.096	0	0	13.599
25	0.292	624	0.292	0.195	0.195	0.098	0	0	13.41
15	0.166	624	0.166	0.162	0.162	0.232	0	0	10.446
10	0.153	622	0.153	0.152	0.152	0.162	0	0	8.823

Table 2. DUC2003 performance, sorted by AVGLAC, averaged length-adjusted score.

7 Future Work

The average length of the headlines produced by our system is well below the length restriction. An improvement in overall coverage will occur with finer toning the parameters that set the size limit on phrase cluster/window and appending additional topics to partial summarized headlines.

When looking at the newly summarized headlines closely, especially with the cases where the participating party of action is missing, a possible next step to advance the system is to embed named-entity identification into the generation step.

We look forward to realizing these hypotheses and contributing improvements to the field of headline summarization.

Acknowledgement

We want to thank Dr. Chin-Yew Lin from ISI for the Yahoo Full Coverage Collection download and many constructive suggestions.

References

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA., pages 68-73. ACM Press.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *proceedings of the 5th*

Conference on Applied Natural Language Processing, Washington, D.C., March 1997.

Rong Lin and Alexander Hauptmann. 2001. Headline generation using a training corpus. In *Second International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, February 18 to 24, 2001. Published by Springer, Lecture Notes in AI.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2002. Automatic headline generation for newspaper stories. In *Proceedings of the Workshop on Text Summarization* post-conference workshop of ACL-02, Philadelphia, PA, 2002.