Template-Filtered Headline Summarization

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

Abstract

Headline summarization is a difficult task because it requires maximizing text content in short summary length while maintaining grammaticality. This paper describes our first attempt toward solving this problem with a system that generates key headline clusters and fine-tunes them using templates.

1 Introduction

Producing headline-length summaries is a challenging summarization problem. Every word becomes important. But the need for grammaticality—or at least intelligibility— sometimes requires the inclusion of non-content words. Forgoing grammaticality, one might compose a "headline" summary by simply listing the most important noun phrases one after another. At the other extreme, one might pick just one fairly indicative sentence of appropriate length, ignoring all other material. Ideally, we want to find a balance between including raw information and supporting intelligibility.

We experimented with methods that integrate content-based and form-based criteria. The process consists two phases. The keyword-clustering component finds headline phrases in the beginning of the text using a list of globally selected keywords. The template filter then uses a collection of pre-specified headline templates and subsequently populates them with headline phrases to produce the resulting headline.

In this paper, we describe in Section 2 previous work. Section 3 describes a study on the use of headline templates. A discussion on the process of selecting and expanding key headline phrases is in Section 4. And Section 5 goes back to the idea of templates but with the help of headline phrases. Future work is discussed in Section 6.

2 Related Work

Several previous systems were developed to address the need for headline-style summaries.

A lossy summarizer that 'translates' news stories into target summaries using the 'IBM-style' statistical machine translation (MT) model was shown in (Banko, et al., 2000). Conditional probabilities for a limited vocabulary and bigram transition probabilities as headline syntax approximation were incorporated into the translation model. It was shown to have worked surprisingly well with a stand-alone evaluation of quantitative analysis on content coverage. The use of a noisy-channel model and a Viterbi search was shown in another MT-inspired headline summarization system (Zajic, et al., 2002). The method was automatically evaluated by BiLingual Evaluation Understudy (Bleu) (Papineni, et al., 2001) and scored 0.1886 with its limited length model.

A nonstatistical system, coupled with linguistically motivated heuristics, using a parse-and-trim approach based on parse trees was reported in (Dorr, et al., 2003). It achieved 0.1341 on Bleu with an average of 8.5 words.

Even though human evaluations were conducted in the past, we still do not have sufficient material to perform a comprehensive comparative evaluation on a large enough scale to claim that one method is superior to others.

3 First Look at the Headline Templates

It is difficult to formulate a rule set that defines how headlines are written. However, we may discover how headlines are related to the templates

derived from them using a training set of 60933 (*headline*, *text*) pairs.

3.1 Template Creation

We view each headline in our training corpus as a potential template. For any new text(s), if we can select an appropriate template from the set and fill it with content words, then we will have a well-structured headline. An abstract representation of the templates suitable for matching against new material is required. In our current work, we build templates at the part-of-speech (POS) level.

3.2 Sequential Recognition of Templates

We tested how well headline templates overlap with the opening sentences of texts by matching POS tags sequentially. The second column of Table 1 shows the percentage of files whose POS-level headline words appeared sequentially within the context described in the first column.

Text Size	Files from corpus (%)
First sentence	20.01
First two sentences	32.41
First three sentences	41.90
All sentences	75.55

Table 1: Study on sequential template matching of a headline against its text, on training data

3.3 Filling Templates with Key Words

Filling POS templates sequentially using tagging information alone is obviously not the most appropriate way to demonstrate the concept of headline summarization using template abstraction, since it completely ignores the semantic information carried by words themselves.

Therefore, using the same set of POS headline templates, we modified the filling procedure. Given a new text, each word (not a stop word) is categorized by its POS tag and ranked within each POS category according to its tf.idf weight. A word with the highest tf.dif weight from that POS category is chosen to fill each placeholder in a template. If the same tag appears more than once in the template, a subsequent placeholder is filled with a word whose weight is the next highest from the same tag category. The score for each filled template is calculated as follows:

$$score\_t(i) = \frac{\sum_{j=1}^{N} W_j}{|desired\_len - template\_len| + 1}$$

where *score_t(i)* denotes the final score assigned to template *i* of up to *N* placeholders and W_j is the tf.idf weight of the word assigned to a placeholder in the template. This scoring mechanism prefers templates with the most desirable length. The highest scoring template-filled headline is chosen as the result.

4 Key Phrase Selection

The headlines generated in Section 3 are grammatical (by virtue of the templates) and reflect some content (by virtue of the tf.idf scores). But there is no guarantee of semantic accuracy! This led us to the search of key phrases as the candidates for filling headline templates. Headline phrases should be expanded from single seed words that are important and uniquely reflect the contents of the text itself. To select the best seed words for key phrase expansion, we studied several keyword selection models, described below.

4. 1 Model Selection

Bag-of-Words Models

1) Sentence Position Model: Sentence position information has long proven useful in identifying topics of texts (Edmundson, 1969). We believe this idea also applies to the selection of headline words. Given a sentence with its position in text, what is the likelihood that it would contain the first appearance of a headline word:

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

$$P(Pos_i) = \frac{Count\_Pos_i}{\sum_{i=1}^{Q} Count\_Pos_Q}$$

Over all *M* texts in the collection and over all words from the corresponding *M* headlines (each has up to *N* words), *Count_Pos* records the number of times that sentence position *i* has the first appearance of any headline word W_j. $P(H_k \mid W_j)$ is a binary feature. This is computed for all sentence positions from *1* to *Q*. Resulting $P(Pos_i)$ is a table on the tendency of each sentence position contain-

ing one or more headlines words (without indicating exact words).

2) Headline Word Position Model: For each headline word W_h, it would most likely first appear at sentence position Pos_i:

$$P(Pos_i \mid W_h) = \frac{Count(Pos_i, W_h)}{\sum_{i=1}^{Q} Count(Pos_Q, W_h)}$$

The difference between models 1 and 2 is that for the sentence position model, statistics were collected for each sentence position *i*; for the headline word position model, information was collected for each headline word W_h.

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

$$P(H_w \mid T_w) = \frac{\sum_{j=1}^{M} (doc\_tf(w,j) \times title\_tf(w,j))}{\sum_{j=1}^{M} doc\_tf(w,j)}$$

doc_tf(w,j) denotes the term frequency of word *w* in the j^{th} document of all *M* documents in the collection. *title_tf(w,j)* is the term frequency of word *w* in the j^{th} title. H_w and T_w are words that appear in both the headline and the text body. For each instance of H_w and T_w pair, $H_w = T_w$.

4) Unigram Headline Model: Unigram probabilities on the headline words from the training set.

5) Bigram Headline Model: Bigram probabilities on the headline words from the training set.

Choice on Model Combinations

Having these five models, we needed to determine which model or model combination is best suited for headline word selection. The blind data was the DUC2001 test set of 108 texts. The reference headlines are the original headlines with a total of 808 words (not including stop words). The evaluation was based on the cumulative unigram overlap between the *n* top-scoring words and the reference headlines. The models are numbered as in Section 4.1. Table 2 shows the effectiveness of each model/model combination on the top 10, 20, 30, 40, and 50 scoring words.

Clearly, for all lengths greater than 10, sentence position (model 1) plays the most important role in selecting headline words. Selecting the top 50 words solely based on position information means that sentences in the beginning of a text are the most informative. However, when we are working with a more restricted length requirement, text model (model 3) adds advantage to the position model (highlighted, 7^{th} from the bottom of Table 2). As a result, the following combination of sentence position and text model was used:

Model(s)	10w	20w	30w	40w	50w
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	139	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 2: Results on model combinations

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

4.2 Phrase Candidates to Fill Templates

Section 4.1 explained how we select headline-worthy words. We now need to expand them into phrases as candidates for filling templates. As illustrated in Table 2 and stated in (Zajic et al., 2002), headlines from newspaper texts mostly use words from the beginning of the text. Therefore, we search for n-gram phrases comprising keywords in the first part of the story. Using the model combination selected in Section 4.1, 10 top-scoring words over the whole story are selected and highlighted in the first 50 words of the text. The system should have the ability of pulling out the largest window of top-scoring words to form the headline. To help achieve grammaticality, we produced bigrams surrounding each headline-worthy word (underlined), as shown in Figure 1. From connecting overlapping bigrams in

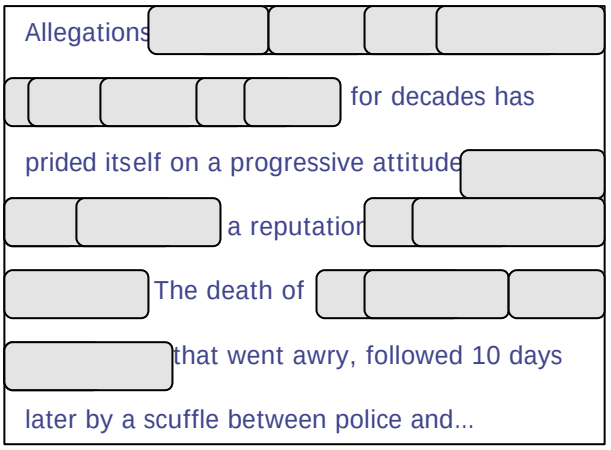

Figure 1: Surrounding bigrams for top-scoring words

sequence, one sees interpretable clusters of words forming. Multiple headline phrases are considered as candidates for template filling. Using a set of hand-written rules, dangling words were removed from the beginning and end of each headline phrase.

5 Filling Templates with Phrases

5.1 Method

Key phrase clustering preserves text content, but lacks the complete and correct representation for structuring phrases. The phrases need to go through a grammar filter/reconstruction stage to gain grammaticality.

A set of headline-worthy phrases with their corresponding POS tags is presented to the template filter. All templates in the collection are matched against each candidate headline phrase. Strict tag matching produces a small number of matching templates. To circumvent this problem, a more general tag-matching criterion, where tags belonging to the same part-of-speech category can be matched interchangeably, was used.

Headline phrases tend to be longer than most of the templates in the collection. This results in only partial matches between the phrases and the templates. A score of fullness on the phrase-template match is computed for each candidate template ft_i:

$$ft_i = \frac{length(t_i) + matched\_length(h_i)}{length(t_i) + length(h_i)}$$

t_i is a candidate template and h_i is a headline phrase. The top-scoring template is used to filter each headline phrase in composing the final multi-phrase headline. Table 3 shows a random selection of the results produced by the system.

Generated Headlines
First Palestinian airlines flight depart Gaza's airport
Jerusalem/ suicide bombers targeted market Friday setting blasts
U.S. Senate outcome apparently rests small undecided voters.
Brussels April 30 European parliament approved Thursday join currency mechanism
Hong Kong strong winds Sunday killing 150 / Philippines leaving hundreds thousands homeless
Chileans wish forget years politics repression

Table 3: System-generated headlines. A headline can be concatenated from several phrases, separated by '/'s

5.2 Evaluation

Ideally, the evaluation should show the system's performance on both content selection and grammaticality. However, it is hard to measure the level of grammaticality achieved by a system computationally. Similar to (Banko, et al., 2000), we restricted the evaluation to a quantitative analysis on content only.

Our system was evaluated on previously unseen DUC2003 test data of 615 files. For each file, headlines generated at various lengths were compared against i) the original headline, and ii) headlines written by four DUC2003 human assessors. The performance metric was to count term overlaps between the generated headlines and the test standards.

Table 4 shows the human agreement and the performance of the system comparing with the two test standards. P and R are the precision and recall scores.

	Assessors'		Generated		
Original	P	R	Length (words)	P	R
	0.3429	0.2336	9	0.1167	0.1566
			12	0.1073	0.2092
			13	0.1075	0.2298
Assessors'	0.2186	0.2186	9	0.1482	0.1351
			12	0.1365	0.1811
			13	0.1368	0.1992

Table 4: Results evaluated using unigram overlap

The system-generated headlines were also evaluated using the automatic summarization evaluation tool ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy,

	Human	Generated
Unigrams	0.292	0.169
Bigrams	0.084	0.042
Trigrams	0.030	0.010
4-grams	0.012	0.002

Table 5: Performance on ROUGE

2003). The ROUGE score is a measure of n-gram recall between candidate headlines and a set of reference headlines. Its simplicity and reliability are gaining audience and becoming a standard for performing automatic comparative summarization evaluation. Table 5 shows the ROUGE performance results for generated headlines with length 12 against headlines written by human assessors.

6 Conclusion and Future Work

Generating summaries with headline-length restriction is hard because of the difficulty of squeezing a full text into a few words in a readable fashion. In practice, it often happens in order to achieve the optimal informativeness, grammatical structure is overlooked, and vice versa. In this paper, we have described a system that was designed to use two methods, individually had exhibited exactly one of the two types of unbalances, and integrated them to yield content and grammaticality.

Structural abstraction at the POS level is shown to be helpful in our current experiment. However, part-of-speech tags do not generalize well and fail to model issues like subcategorization and other lexical semantic effects. This problem was seen from the fact that there are half as many templates as the original headlines. A more refined pattern language, for example taking into account named entity types and verb clusters, will further improve performance. We intend to incorporate additional natural language processing tools to create a more sophisticated and richer hierarchical structure for headline summarization.

References

Michele Banko, Vibhu Mittal, and Michael Witbrock. 2000. Headline generation based on statistical translation. In *ACL-2000*, pp. 318-325.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: a parse-and-trim approach to headline generation. In *Proceedings of Workshop on Automatic Summarization*, 2003.

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

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL 2003*, pp.150–157.

Rong Lin and Alexander Hauptmann. 2001. Headline generation using a training corpus. *In CICLING 2000.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jin Zhu. 2001. IBM research report Bleu: a method for automatic evaluation of machine translation. In *IBM Research Division Technical Report*, RC22176 (W0109-22).

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