Balanced Coverage of Aspects for Text Summarization

Takuya Makino
Interdisciplinary Graduate
School of Science and Engineering
Tokyo Institute of Technology
makino@lr.pi.titech.ac.jp

Hiroya Takamura **Manabu Okumura**
Precision and Intelligence Laboratory,
Tokyo Institute of Technology
{takamura ,oku}@pi.titech.ac.jp

Abstract

In this paper, we present a new text summarization model based on max-min problem to cover aspects. Aspects are pre-defined for each category of document cluster; for example, the category *Accidents and Natural disasters* are *WHY* and *DAMAGES*. Our goal is to generate a summary that covers these aspects. In order to calculate the score indicating the aspect coverage, we use the maximum-entropy classifier that predicts whether each sentence reflects the aspect or not. In our model, the score indicating the coverage for each aspect is calculated and the minimum of the scores of the aspects is going to be maximized so that the summary contains all the aspects. Through the summarization experiments on the TAC dataset, we show that our model outperforms a state-of-the-art summarization system in terms of ROUGE-2.

1 Introduction

Automatic text summarization task is to create a summary, or a short concise document that describes the content of a given set of documents (Mani, 2001). Text summarization technique is an important when there is a large amount of text about the event that the reader wants to know about. Text Analysis Conferences (TAC) has introduced *the guided summarization task* that is different from generic summarization task nor from query-focused summarization task. In the guided summarization task, the summary should contain the aspect oriented information for all aspects.

To cover aspects, Vasudeva (2010) proposed an extractive approach that selects a sentence greedily, which reflects a certain aspect most strongly. However, this approach may solve the problem only locally. Thus, we formulate the text summarization task as a combination of a maximum coverage problem with knapsack constraint (MCKP) and max-min problem, which can be represented as an Integer Linear Programming Problem (ILP). In our model, the score indicating the coverage for each aspect is calculated and the minimum of the scores of the aspects is going to be maximized so that the summary contains all the aspects.

In the maximum-coverage summarization models, the frequency of the word in the document and the position of the word in the document are often used to determine the importance of the word (or conceptual unit). In the guided summarization task , however, a summary should contain the aspect-oriented information. Therefore, the frequency and the position of a word is sometimes not sufficient for determining its importance. For example, when the topic of a document cluster is *Accidents and Natural disasters* and one of aspects is *COUNTER-MEASURES*, which corresponds to rescue efforts. If few documents describe rescue efforts, a simple maximum-coverage model would fail to cover this aspect. We therefore propose a model which is based not solely on the maximum coverage problem, but also on the max-min problem, in which the balanced coverage of aspects is implemented.

In order to calculate the score indicating the aspect coverage, we use the maximum-entropy classifier that predicts whether each sentence reflects the

aspect or not. Although the maximum-entropy classifier is a classifier, we use the output probability which is actually the conditional probability that the aspect is reflected by the sentence. In the training of the maximum-entropy classifier, we need a training dataset. Since it is sometimes impractical to assume the availability of the labeled training data or the availability of the sufficient labelling workload by human annotators, we reduce the labelling workload by letting the annotators annotate section names of Wikipedia articles of the similar category, instead of each sentence, with aspect labels. For example, *Cause* is labeled with *WHY*. We use the sentences in the labeled sections as the labeled training data. Although the labels in the data can be noisy, we believe that it still works as a training dataset. Furthermore, we can benefit from the huge amount of data in Wikipedia.

2 Related Work

2.1 Integer Linear Programming Problem

Goldstein (2000) used sequential sentence selection in combination with maximal marginal relevance (MMR), which gives penalty to sentences that are similar to the already selected sentences. Since their method is a greedy procedure of selecting sentences and does not measure the goodness of the entire summary, global summarization models based on ILP have recently been studied intensively. ILP is a kind of linear programming problems, in which the values of the variables are constrained to integers. McDonald (2007) formulated the text summarization task as an ILP and applied an approximate dynamic programming decoding. Takamura (2009) formulated the text summarization task as an augmented maximum coverage problem, whose objective function is a combination of coverage and relevance to the subject of document cluster.

2.2 Summarizatin Approach Using Wikipedia

There are some summarization approaches that use Wikipedia. Vasudeva (2010) predicted the score indicating the aspect for each sentence, and selected a sentence that has the highest score for each aspect. Fujii (2009) automatically determined aspects of search term using sections in Wikipedia, and extracted a sentence that most highly represents the aspect with Support Vector Machine (SVM).

3 Proposed Approach

3.1 Prediction of the Score of the Aspect for each Sentence

The aspect is important information to understand the specific content of a document cluster. Aspects are predefined for each category; for example, the aspects of the topic *Accidents and Natural disasters* are *WHAT*, *WHEN*, *WHERE*, *WHY*, *WHO AFFECTED*, *DAMAGES*, *DAMAGES* and *COUNTERMEASURES*. Descriptions of aspects are as follows.

WHY: reasons for accident/disaster
WHO AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the accident/disaster
DAMAGES: damages caused by the accident/disaster
COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the accident/disaster

In order to calculate the score indicating the aspect coverage, we use the maximum-entropy classifier that predicts whether each sentence reflects the aspect or not. Although the maximum-entropy classifier is a classifier, we use the output probability which is actually the conditional probability that the aspect is reflected by the sentence because our system is extractive. We show the feature value in the equation (1) below:

$$\phi_k(j,y) = \begin{cases} 1 & \text{if } n\text{-}gram\ k \text{ appears in } j \text{ and } y = a, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Let $\phi_k(j,y)$ denote a feature value which is 1 if sentence j contains $n\text{-}gram\ k$ and label y is a, otherwise 0. $n\text{-}gram$ in equation (1) represents unigrams and bigrams in the training data. The conditional probability that the aspect a is reflected by the sentence j is expressed as follows:

$$p(y|j) = \frac{1}{Z(j)} \exp\left(\sum_k \lambda_k \phi_k(j,y)\right), \quad (2)$$

Table 1: The subset of queries to collect articles from Wikipedia

Topic	Query
Accidents and Natural disasters	*Natural disaster*, *accident*
Attacks	*Attack, Terro*
Health and Safety	*Health, Safety*
Endangered Resources	*Resources, energy*
Trials and Investigations (Criminal/Legal/Other)	*Investigations*, *Trials*, *Legal action* *Criminal procedure*, *Case law*

Table 2: Aspects and the subset of section names

Aspect	Section name
WHY	*Causes*, *Cause and results*, *Reasons for crash*, *Probable cause*
WHO AFFECTED	*Victims*, *Fatalities*, *Injuries*
DAMAGES	*Damages*, *Observed damage*, *Collision*
COUNTERMEASURES	*Rescue*, *Emergency response*, *Recovery*

where $Z(j)$ is the normalization factor and λ_k is the weight of feature k.

3.2 Labeling Sections in Wikipedia

Fujii et al., (2009) trained the classifier for sentence extraction through the use of sections in Wikipedia. We also use Wikipedia as training data. We have three benefits for using Wikipedia as the training data: 1)all articles are categorized, 2)each article has sections, 3)Wikipedia has a large amount of articles. We show the outline of processes from creating process of the training data to training the maximum-entropy classifier.

1. First, we collect a set of articles whose categories are similar to a topic in TAC. The set of queries for collecting such articles are determined by a human annotator.

2. Next, the annotator labels section names of Wikipedia articles of the similar categories with aspect labels.

3. Finally, we use sentences in the section names labeled with aspect a as positive instances, and sentences in the section names labeled with aspect not involving a as negative instances.

Since it is sometimes impractical to assume the availability of the labeled training data or the availability is sometimes impractical to assume the availability of the labeled training data or the availability of the sufficient labelling workload by human annotators, we reduce the labelling workload by letting the annotators annotate section names of Wikipedia articles of the similar category, instead of each sentence, with aspect labels. For example, Causes is labeled with *WHY*. We use the sentences in the labeled sections as the positive instances. Although the labels in the data can be noisy, we believe that it still works as a training dataset. Furthermore, we can benefit from the huge amount of data in Wikipedia. We show aspects we considered in Table 2. Although there are some more aspects: *WHAT*, *WHEN* and *WHERE* in TAC, we did not use them, because aspect *WHAT* can probably be covered by the maximum coverage part of our model, and aspects *WHEN* and *WHERE* would be difficult to label.

3.3 Modeling for Balanced Coverage of Aspects

Takamura et al. (2008) modeled text summarization as the combination of coverage and relevance. In fact, their model is based on the intuition that the summary should not contain redundant contents, but the sentences to be selected have to be relevant to the main content of the document cluster. We integrated the new term into the objective function of the previous work to cover aspects. Our model is formalized as below:

$$\max \quad (1-\beta)\left\{\alpha\sum_i w_i c_i + (1-\alpha)\sum_j\left(\sum_i w_i o_{ij}\right)s_j\right\} + \beta z, \quad (3)$$

subject to

$$\forall j, \sum_j s_j o_{ij} \geq c_i,$$
$$\forall i, \forall j, s_j o_{ij} \leq c_i,$$
$$\sum_j s_j l_j \leq L,$$
$$\forall a \in aspects, \sum_j s_j p_{aj} \geq z,$$
$$\forall i, c_i \in \{0,1\},$$
$$\forall j, s_j \in \{0,1\}.$$

Our model is going to maximize the objective function under some constraints. Here, let c_i denote 1 if the summary contains conceptual unit i, otherwise 0, s_j denote 1 if the summary contains conceptual unit j, otherwise 0, o_{ij} denote 1 if the sentence j contains conceptual unit i, and w_i denote the weight of conceptual unit i. The first term and the second term in equation (3) correspond to coverage and relevance with subject of the document cluster respectively. These two terms are linearly combined by parameter α. Larger value of α indicates more weight on coverage, rather than on relevance. The third term is the proposed term to cover aspects. z denotes the minimum of the scores of aspects in the summary. The score of the aspect a in the summary is expressed as the summation of the scores of the aspect a for sentence j in the summary; $\sum_j s_j p_{aj}$.

Thus, maximizing z means that the summary contains all the aspects. The summary length is at most L and l_j is the length of the sentence j where length means the number of words. We use only top N scores for each aspect.

We linearly combined the original objective function and the proposed term corresponding to the aspect coverage by β. Larger value of β indicates more weight on the aspect coverage. When $\beta = 0$, our model is reduced to the model of Takamura et al. (2009).

3.4 Updating the Summary

For the update summarization task, summarization of B set in TAC, we implemented two approaches. One approach penalizes the weight of a conceptual unit when the conceptual unit already appeared in A set, while another approach does not.

We describe the former approach. We define the new weight w_i' of the conceptual unit i in B set as follows:

$$w_i' = \begin{cases} w_i - \epsilon w_{ia} & \text{if } i \text{ already appeared in A set,} \\ w_i & \text{otherwise,} \end{cases} \quad (4)$$

where w_{ia} is the weight of the conceptual unit i in A set and ϵ is a positive constant.

4 Experiments

4.1 Preprocessing

We used WP2TXT[1] to parse the XML file [3] distributed from Wikipedia. In the training of the maximum-entropy classifier, we identified named entity in the training data with Illinoi Named Entity Tagger[2] and masked the named entity with its named entity tag because we do not need named entities. We show an example of this masking process:

Mrs. Morely died while in Lima.
→ PER died while in LOC.

Furthermore, we did not use short sentences as training data, because very short sentences tend to be erroneous ones. We only use sentences whose length is larger than 5 words.

We used bigrams unit as conceptual units. Following Yih (2007), the weight of the conceptual unit is the average of the binary value indicating whether the conditional unit appears in the document or not and the binary value indicating whether the conceptual unit appears within the first 100 words in the document or not. All words in the document cluster are lemmatized.

The scores of aspects are normalized so that the sum of their squares is 1. We used top 10 (N =

[1] http://wp2txt.rubyforge.org/

[3] http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

[2] http://cogcomp.cs.illinois.edu/page/software_view/4

Table 3: ROUGE on A set

run id	ROUGE-2	ROUGE-SU4
39	0.1188 (0.093 ± 0.029)	0.1479 (0.127 ± 0.033)
40	0.1188 (0.093 ± 0.029)	0.1479 (0.127 ± 0.033)

Table 4: ROUGE on B set

run id	ROUGE-2	ROUGE-SU4
39	0.085 (0.070 ± 0.014)	0.1172 (0.1094 ± 0.014)
40	0.080 (0.070 ± 0.014)	0.1151 (0.1094 ± 0.014)

Table 5: average Pyramid score of 2 runs on A set

topic	average Pyramid score
Accidents and Natural disasters	0.569 (0.474 ± 0.131)
Attacks	0.480 (0.424 ± 0.122)
Health and Safety	0.296 (0.282 ± 0.078)
Endangered resources	0.252 (0.285 ± 0.078)
Investigations and Trials (Criminal/Legal/Other)	0.545 (0.397 ± 0.116)
average	0.427 (0.372 ± 0.098)

10) scores for each aspect when we solve the text summarization problem. We set parameters α and β to the values that maximize ROUGE-2 for the document clusters of the same topic in TAC 2010 dataset.

4.2 Dataset

In TAC 2011, all participants are supposed to summarize 46 document clusters. Each document cluster consists of 10 news articles and is categorized into only one topic. The topic of each cluster is given beforehand. The summary length is at most 100 ($L = 100$) words. For each topic, aspects are pre-defined.

4.3 Result

The indices of our systems are 39 and 40 in TAC 2011; system 39 does not penalize the weight of a conceptual unit when the conceptual unit already appeared in A set, while system 40 does. Thus, the two systems generate exactly the same summary for A set.

We show the ROUGE scores on A set and B set respectively in Table 3 and Table 4. The average and the standard derivation of ROUGE scores of all the TAC participants are showed in the parentheses. Our systems worked well for A set in terms of ROUGE-2 and ROUGE-SU4 (Table 3). For B set, our systems showed less improvement over the average score than for A set (Table 4). A rather simplistic approach for update summarization based on the weight reduction of the conceptual units failed to give a good model. We would need a more sophisticated approach for update summarization. Next, we show the Pyramid scores for A set and B set respectively in Table 5 and Table 6, where the values are the average of the Pyramid scores for each aspect. At the bottom of each table, we added the average of those average Pyramid scores. Table 4 shows the average Pyramid scores of two systems 39 and 40. Note that although the two systems generate exactly the same summaries for A set, their Pyramid scores can be different from each other because the scores are given by human evaluators. Except for topic Endangered resources, our pyramid scores are good for each topic for A set (Table 5), but not for B set (Table 6).

We first examine the informative features in the topic Accidents and Natural disasters. The SCU "*to prevent the oil from reaching the shoreline*" in D1124E-A is labeled as *COUNTERMEASURES* by annotators and our summary contains this SCU. In the maximum entropy classifier for aspects *COUNTERMEASURES*, this feature "*prevent*" was given the weight in the top $25[\%]$ and enables our system to cover this SCU. The SCU "*study by British experts said the eruption was most likely caused by drilling for gas*" is labeled as *WHY*. Contrary to our expectation, the features "*caused*" and "*caused by*", which are supposedly cue phrases to extract the answers to why-question, are not given a high weight in the maximum entropy classifier for aspect *WHY*. Instead, the features "*weather*", "*wind*" and "*crash* are more weighted. Words "*weather*" and "*wind*" refer to causal events, not cue phrases. This is probably because cue phrases are used also for other aspects, and not given a high weight. Word "*crash*" refers to the effect, rather than to the cause. The reason of the high weight on this word would be that the cause and the effect appears often together in the same sentence. However, our system failed to cover aspects for some topics. Thus we refined the training data and conducted additional experiments

Table 6: Pyramid scores of runs on B set

topic	average Pyramid score on 39	average Pyramid score on 40
Accidents and Natural disasters	0.312 (0.279 ± 0.087)	0.299 (0.279 ± 0.087)
Attacks	0.364 (0.298 ± 0.073)	0.353 (0.298 ± 0.073)
Health and Safety	0.198 (0.223 ± 0.059)	0.228 (0.223 ± 0.059)
Endangered resources	0.225 (0.289 ± 0.084)	0.221 (0.289 ± 0.084)
Investigations and Trials (Criminal/Legal/Other)	0.264 (0.275 ± 0.075)	0.264 (0.275 ± 0.075)
average	0.272 (0.271 ± 0.060)	0.273 (0.271 ± 0.060)

Table 7: informative features for *WHY*

feature	the contributor of SCU labeled as *WHY* in the summary
wind	*Loose tow ropes, which broke in high wind and waves, probably caused the collision (D115C-A)*
failure	*plane's failure to fly at 35,000 feet (D1105A-A)*
crash	*The spill occurred when a barge carrying a crane crashed into the tanker (D1115C-A)*
weather	*Officials say bad weather was the cause of the accident (D1130F-A)*

$< topicid =$ "$D1118D$" $category =$ "$\underline{5}$"$>$

$< title > HawkinsRobertVanMaur < /title >$
...

Figure 1: GuidedSumm_topics.xml

...
$<scuuid$="2" $label$="$Ashootingspree(\underline{2}.1)$"$>$
$<contributorlabel$="$ashootingspree$"$>$
$<partlabel$="$ashootingspree$"$start$="162"
end="178"$/>$
$</contributor>$
$<contributorlabel$="$openedfire$"$>$
$<partlabel$="$openedfire$"$start$="773"end="784"$/ >$
...

Figure 2: D1118-A-ADEF.pyr

in Section 5.

Note that the evaluation on the document cluster D1118D may not be valid, since the document cluster is categorized into Trials and Investigations in Guidedsumm_topics.xml, but into Attacks in D1180DA.pyr (Figure 1 and Figure 2).

5 Additional Experiment

To improve the quality of the summary, we refined the training data. We modified the queries used to collect articles of the similar topics from Wikipedia (Table 8).

The baseline system is the model of Takamura et al. (2009), which does not take the coverage of aspects into account.

We also removed punctuation marks when we created the conceptual units, so that the bigrams representing conceptual units do not contain any punctuation marks.

We set parameters α and β to the values that maximize ROUGE-2 for the document clusters of the same topic in TAC 2010 dataset.

Our modified system improved ROUGE-2 (Table 9) over our original system officially submitted to TAC 2011 (Table 3). However, we obtained only insignificant improvement over the baseline system.

6 Conclusion

We proposed a summarization approach that explicitly takes aspects into consideration. Although our system showed a good performance in the guided summarization task, it showed only insignificant improvement over the simple maximum coverage model by Takamura et al. (2009), which does not take aspects into consideration. One possible reason of the insignificance would be the low quality of the training data for the maximum entropy classifier of aspects. We plan to use TAC data annotated with

Table 8: The subset of sophisticated queries to collect articles from Wikipedia

Topic	Query
Accidents and Natural disasters	*natural disaster, accident*
Attacks	*Spree shooting, Massacre, Suicide bombing, Hijacking*
Health and Safety	*HIV/AIDS, Organ transplant, Food recall, Health disaster*
Endangered Resources	*Endangered species, Endangered animals, Water pollution*
Trials and Investigations (Criminal/Legal/Other)	*Murder trial, Robbery trial, Trials in, manslaughter, sex crime trial*

Table 9: ROUGE on A set in the additional experiment

topic	baseline	aspect-coverage
Accidents and Natural disaster	0.1383	0.1383
Attacks	0.1568	0.1568
Health and Safety	0.1143	0.1148
Endangered resources	0.0826	0.0843
Trials and Investigations (Criminal/Legal/Others)	0.1234	0.1234
micro average	0.1238	0.1243
macro average	0.1231	0.1236

aspect labels as training data.

References

Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki: Multi-Document Summarization by Maximizing Informative Content-Words. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp.1776–1782 (2007).

Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharath Reddy, Sudheer Kovelamudi, Srikanth Reddy Vaddepally, Radheshyam Nanduri, Kiran Kumar N, Santhosh Gsk, Prasad Pingali: IIIT Hyderabad in Guided Summarization and Knowledge Base Population. TAC (2010).

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

Ryan McDonald: A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European conference on IR reseach*, ECIR'07, pp.557–564, Berlin, Hidelberg, Springer-Verlag (2007).

Dan Gillick and Benoit Favre: A Scalable Global Model for Summarization. In *Proceedings of NAACL Workshop on Integer Linear Programming for Natural Language Processing* (2009).

Hiroya Takamura and Manabu Okumura: Text Summarization Model based on Maximum Coverage Problem and its Variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 781–789. Association for Computational Linguistics (2009).

Atsushi Fujii and Akihiko Sanjouba: Modeling Term Descriptions Using Wikipedia and its Application to Encyclopedic Search. Japanese Society for Artificial Intelligence, SIG-SWO-A803-01, pp.1–8 (2009).