Summarization using Wikipedia

Shu Gong*, Youli Qu, Shengfeng Tian
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
Email: monicashu452@gmail.com, ylqu@bjtu.edu.cn, sftian@bjtu.edu.cn

***Abstract*—Summarization utilizes the processing speed of computer to provide a concise and comprehensive glance at the data mass. This paper presents our work in the Text Analysis Conference 2010 guided summarization task. We make use of Wikipedia, currently the world's largest online encyclopedia, to catch the semantic meaning under words. Three features are extracted for sentence selection, based on disambiguated Wikipedia concepts and their first paragraphs in the corresponding Wikipedia articles. Our system ranks in the middle tier of the 43 peer systems (including 2 baselines). From the analysis of evaluation results, we get the following findings: Firstly, we can find a sufficient number of Wikipedia concepts in each topic. Secondly, it is the context representation but not the number of concepts that affects system performance for a topic. Finally, highly related first paragraphs in Wikipedia article significantly improve system performance.**

I. Introduction

Summarization utilizes the processing speed of computer to provide a concise and comprehensive glance at the data mass. The output of a summarization system, called a summary, is a compressed version of the original documents collection, capturing important and non-redundant information.

The Text Analysis Conference (TAC), previously known as Document Understanding Conferences (DUC), has continued its annual evaluation of summarization since 2000. The whole dataset covers 46 topics, with 20 documents for each topic. This year's topic specification includes a category node instead of the topic description used in the past. In total, there are five categories with detailed aspects. Each set's topic falls into one of the categories. The topic distribution over categories is shown in Table I.

For each topic, its documents are chronologically ordered and divided equally into two sets A and B. The 10 Documents in the set A of a topic are all earlier than those in the set B of the same topic. Systems are required to generate a 100-word summary for each topic set A or B, covering different aspects of the category the set related to. Following the requirement of update summarization, summary for a topic's set B should avoid the repetition of information already found in the topic's set A.

Different from past years, this year's guided summarization task encourages deep semantic analysis of text instead of using only word frequency. Aspect coverage is also a new requirement. Our system tries to use concept analysis to catch semantically important information. We use Wikipedia to extract three concept features for each sentence. These features assess both salient information and aspect coverage in a sentence. We combine these features with other widely used features for sentence ranking and summary sentence selection.

In the following section, we first explain our system in detail, from Wikipedia concept identification, concept weighting, to feature calculation and summary generation. Then we present the evaluation results with analysis of each feature. Finally, we conclude with general remarks about using Wikipedia in summarization, address the promising aspects and the existing problems of our work, and present possible directions for future research.

* This work was done during the first author's internship at HP Labs China.

II. Concept Identification

The documents of TAC 2010 dataset are news collection precollected from several news websites, and are provided in the original html format. Thus, for each topic set $D = \{d_1, d_2, ..., d_T\}$, we have to do html parsing to extract the main text content first.

In order to use the concepts in Wikipedia, we need to find out which Wikipedia concepts are existed in the document set. We use the wikification function provided in WikiMiner [1] for this purpose. Text wikification is defined as "the task of automatically extracting the most important words and phrases in the document, and identifying for each such keyword the appropriate link to a Wikipedia article" [2]. It includes three steps: keyphrase detection, concept disambiguation and link detection. We use the first two steps for concept identification.

A. Keyphrase Detection

In the detection step, keyphrases are identified using the "Keyphraseness" measurement, which indicates the wikified probability of a phrase in Wikipedia.

For each phrase a, its keyphraseness can be calculated as its link probability, which is defined as:

$$Keyphraseness(a) = P_{link}(a) \approx \frac{count(a_{Link})}{count(a)} \quad (1)$$

in which, $count(a_{Link})$ is the number of times that this phrase has been used as the anchor text of wiki-links in all Wikipedia articles, and $count(a)$ is its total occurrence, as wiki-link or normal text. Phrases with small link probability are discarded. We set this threshold to be 0.01.

After detection, a collection of keyphrases $K = \{K_1, K_2, ..., K_M\}$ is found.

B. Concept Disambiguation

In the disambiguation step, each keyphrase's sense used in the current context is found and represented using an article of Wikipedia, which stands for a concept.

Firstly, for a keyphrase which is one of the variants of a concept, we can follow its redirect page in Wikipedia to reach its corresponding concept. Secondly, for a keyphrase which has been linked to only one target, it is unambiguous. The target article is just its corresponding concept. Finally, for an ambiguous keyphrase K_m which has been linked to many different target concepts, three features are used to evaluate each candidate sense s, balancing the global usage and local context semantic.

- "Commonness" has been adopted by many disambiguation works [1], [2], The sense's commonness means how often this sense has been used for a phrase, indicating global usage. It is defined as:

$$Commonness(s) = P(s \mid K_m) = \frac{count(s, K_m)}{LinkCount(K_m)} \tag{2}$$

  in which, $count(s, K_m)$ means the number of times that this sense is used as K_m's target in Wikipedia.
- "Relatedness". "Relatedness" and the next "ContextQuality" feature are both based on an observation: there always exist some unambiguous keyphrases in the context. These keyphrases' corresponding concepts form a context. Sense's relatedness with all context concepts is evaluated to solve the semantic uncertainty of ambiguous keyphrase. Given a set of context concepts $C = \{c_1, c_2, ..., c_N\}$, the sense's relatedness with the context is defined as the weighted relatedness between the sense and each context concept:

$$relatedness(s, C) = \sum_{i=1}^{N} wt(c_i) * relatedness(s, c_i) \tag{3}$$

$$wt(c_i) = \frac{P_{link}(c_i) + 2 * \sum_{j=1}^{N} relatedness(c_i, c_j)}{3} \tag{4}$$

$$relatedness(s, c_i) = \frac{\log(\max(|I_s|, |I_{c_i}|)) - \log(|I_s \cap I_{c_i}|)}{\log(|W|) - \log(\min(|I_s|, |I_{c_i}|))} \tag{5}$$

  in which, I_s and I_{c_i} are the sets of Wikipedia articles that contain wiki-links pointing to s and c_i respectively, W stands for the whole set of Wikipedia articles, and $wt(c_i)$ is the weight of the context concept c_i.
- "ContextQuality" is the total importance of all context concepts, defined as:

$$ContextQuality(s) = \sum_{i=1}^{N} wt(c_i) \tag{6}$$

This three features are used to train a disambiguation classifier using bagged C4.5 algorithm. After disambiguation, all concepts included in the topic set and their corresponding Wikipedia articles are ready for use. We denote this concept set as $KC = \{KC_1, KC_2, ..., KC_K\}$.

Considering the length of news documents varies from a few sentences to many long paragraphs, it's hard to find enough context concepts for short documents. As the documents in each topic set are focused on the same topic, we concatenate their content together to get a full text of the topic set as the input of wikification.

III. Concept Weighting

Every concept is weighted considering its relatedness to the context and other concepts, and also its frequency in the topic set.

$$wt(K_m) = freq(K_m, D) * avgR(K_m, C) * avgR(K_m, K) \tag{7}$$

$$avgR(K_m, C) = \frac{\sum_{i=1}^{N} wt(c_i) * relatedness(K_m, c_i)}{N} \tag{8}$$

$$avgR(K_m, K) = \frac{\sum_{j=1}^{M} relatedness(K_m, K_j)}{M - 1} \tag{9}$$

in which, $avgR(K_m, C)$ is the concept's average weighted relatedness with all unambiguous context term, while $avgR(K_m, K)$ is the concept's average relatedness with all other concepts.

IV. Features Calculation

Based on the weight of concepts, we extract three wiki-related features and two widely used features for each sentence.

Suppose sentence S_q contains a subset of concepts $T \in KC$, we have:

A. WC(Wikipedia Concept)

Feature WC stands for the salient information coverage of a sentence. It's defined as the weighted sum of the concepts which exist in the sentence:

$$WC(S_q) = \sum_{t \in T} freq(t, S_q) * wt(t) \tag{10}$$

B. WNC(Wikipedia New Concept)

Feature WNC stands for the new but also salient information coverage of a sentence. It's defined as the weighted sum of the new concepts which exist in the sentence. New concepts are found by comparing the new occurring concepts of set B with its earlier set A. Suppose the concepts in the earlier set are denoted as OT, new concepts are defined as: $NT = \{nt \mid nt \in T \cap nt \notin OT\}$. then:

$$WNC(S_q) = \sum_{nt \in NT} freq(nt, S_q) * wt(nt) \tag{11}$$

As set A doesn't has an earlier set, this feature only works for set B.

C. WFP(Wikipedia First Paragraph)

Because the aspects are all developed based on the model summaries of past years' summarization tasks, they cover the main information points needed by different categories of topics for generating good summaries. We could follow the attributes of listed aspects to summarize each topic, such as for aspect *WHEN*, sentences containing date are more important. In this case, a robust summarization system needs to do aspect extraction when dealing with random document sets, as aspects vary from category to category.

We try to solve this problem by taking advantage of the broad category coverage of Wikipedia articles, together with the first paragraph of Wikipedia articles which can be seemed as a human written summary. The categories in Wikipedia vary from objects to events, from actions to opinions. All categories of TAC 2010 dataset are included in Wikipedia. From example, the category of "Accidents and Natural Disasters" is an event, and Wikipedia contains a "Natural disasters" category including articles like "Wenchuan earthquake", "Typhoon Nina", "H1N1 Influenza", etc. As the first paragraph of each article is a brief introduction of the concept, it includes all important aspects of the concept's category. Thus, the first paragraphs of concepts can be used to assess sentence's aspect coverage.

As a result, the first paragraphs of all concepts are extracted and broken into a sentence set $FP = f_1, f_2, ...f_P$. WFP is defined as:

$$WFP(S_q) = sim(FP, S_q) * \sum_{t \in T} freq(t, S_q) * wt(t) \quad (12)$$

$$sim(FP, S_q) = \max(\cos(S_q, f_p), p \in [1, P]) \quad (13)$$

D. POS(POSition)

The sentence position feature is calculated following the one used in MEAD [3], which is defined as:

$$POS(S_q) = \frac{1}{\sqrt{id}} \quad (14)$$

in which *id* is the sentence's occurring sequence index in document.

E. LEN(LENgth)

Sentence length is used as a filtering feature. The sentences which are shorter than a minimum length limit are considered as meaningless, thus are filtered out before sentence selection.

V. Summary Generation

The five features above are input into the MEAD system to generate a sentence-based extractive summary. We use the default classifier for linearly feature combination. The default reranker based on sentence similarity is used for redundancy removal.

VI. Evaluation and Analysis

We tune the feature parameters using DUC2007 dataset and choose the combination that leads to the best performance. Except the feature LEN, which are tuned with 3, 5, 7 respectively, the other four features are tuned in 0.1 step from 0 to 1, and summed up to 1. Feature parameters used for the submit version for set A and B are listed out in Table II.

TABLE II
Feature parameters used for submitted summaries

Set	LEN	POS	WC	WFP	WNC
A	5	0.1	0.8	0.1	–
B	7	0.1	0	0.8	0.1

Traditional, TAC 2010 evaluates system generated summaries against the model summaries written by human summarizers using the following three of metrics: 1) ROUGE scores based on overlapping units [4], 2) BE score based on basic elements matching [5], 3) Manual scores for Summary Content Units(SCUs) matching [6], linguistic quality and responsiveness. Other than the above metrics, two additional manual evaluations are added this year for category and aspect coverage.

Our system's id in TAC 2010 is 32. Firstly, we show the system's ranks of all metrics among the 43 peers in Table III. The system gets similar performance in all evaluations.

TABLE III
System ranks of different evaluation metrics in TAC2010

Metrics		Set-A	Set-B
ROUGE	R-2	25	28
	R-SU4	27	31
BE		27	30
Manual	Avg modified(pyramid) score	24	25
	Avg SCUs num	25	26
	Avg repetitions num	19	8
	Macro avg modified score with 3 models	24	25
	Avg linguistic quality	25	19
	Avg overall responsiveness	24	23
Manual (Aspect)	Avg modified(pyramid) score	26	27
	Macro avg modified score with 3 models	26	27
Manual (Category)	Avg modified(pyramid) score	25	24
	Avg SCUs num	26	26
	Avg repetitions num	19	8
	Macro avg modified score with 3 models	25	24
	Avg linguistic quality	26	21
	Avg overall responsiveness	24	22

Then, in order to get a detailed view of the system performance for each topic, we extract all peers' ROUGE scores of all topics in set A and B. We convert the rank r of our system among other peers into $1/(1 + log(r))$ to indicate its performance among peers, while 1 means the best and the smaller the worse. As shown in Fig. 1.

For set A, which corresponding to a multi-documents summarization task, topic 30 and 37 have got a better performance

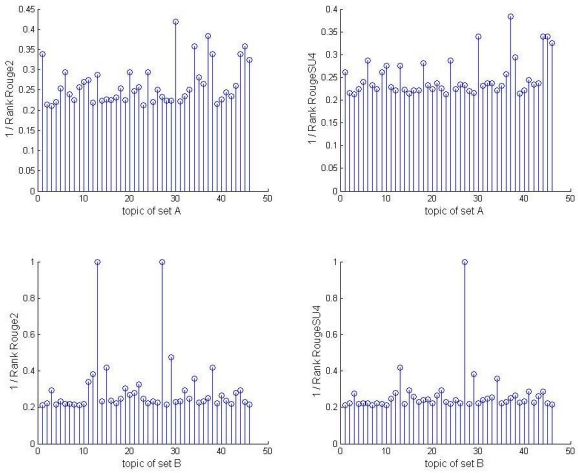

Fig. 1. System performance for each topic compared to other peers of TAC 2010

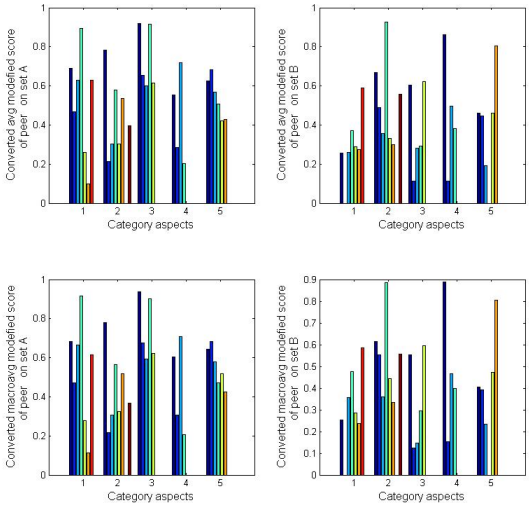

Fig. 2. System performance for each aspect compared to other peers of TAC 2010 on average modified (pyramid) score. The aspects are grouped in category. The other metric "macroaverage modified score with 3 models", which hasn't been list here, gets similar results.

compared to other peers, followed by topic 44, 45 and 46. Topic 2, 3, 23 and 39 have shown very poor performance. For set B, which corresponding to an updating summarization task, topic 27 has got the highest rank among all peers for both measures, followed by topic 13, 29. Topic 28 and 46 got the worst performance.

Finally we present the detailed view of our system's performance for aspects and category. We transform the score of each metric into a value between 0 and 1 by dividing our score by the max score of all peers.

Overall, the performance for category (Fig. 3) is better than on aspect (Fig. 2). While for set A, the system performs best on category 3(Health and Safety), followed by category 1(Accidents and Natural Disasters) and 5(Criminal/Legal/Other Investigations and Trials). For set B, the best performance is on category 2(Criminal/Terrorist Attacks), followed by category 5 and 4(Health and Safety).

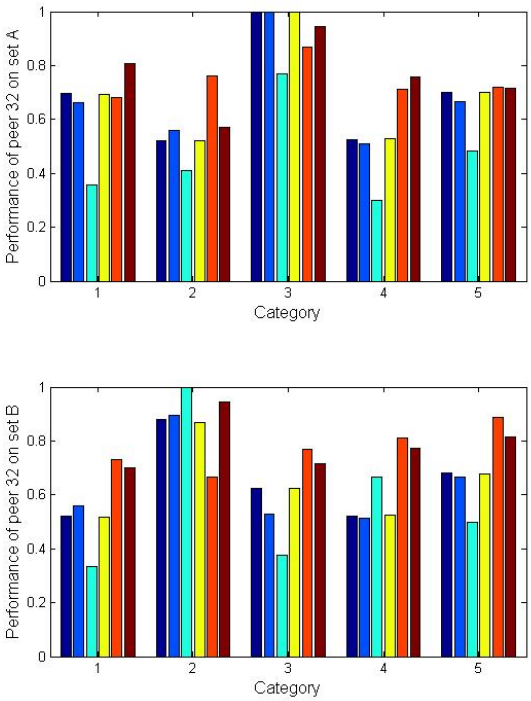

Fig. 3. System performance for each topic compared to other peers of TAC 2010. Each group stands for a category, while each bar in the group stands for one of the six metrics used in traditional manual evaluation(See Table III).

To summarize, the system get better performance for the multi-documents summarization task, while get the highest rank in the updating task on some topics.

In the following, we analyze our system's performance of each feature mainly based on the ROUGE-2 metric, as it's the most widely used one, and is also more quantized. ROUGE-SU4 shows a good correlation with ROUGE-2, thus we don't repeatedly list out the analysis based on it.

A. *Wikified concepts*

As for set A, the feature WC takes major responsibility of the performance with weight 0.8, we analyze the influence of concept number and concept quality of each topic's set A.

1) Concept number: We first calculate the proportions of concept in topic vocabulary and topic length, which is the total number of tokens shown in a topic. From Fig. 4 we can see that the system's performance on ROUGE scores hasn't been affected by how many concepts we can find.

2) Concept quality: Then we look into the concepts we have found. In Table IV, we list out some topics with their ranks in ROUGE-2. A short description is written for each topic to provide a simple idea of what it talks about. From

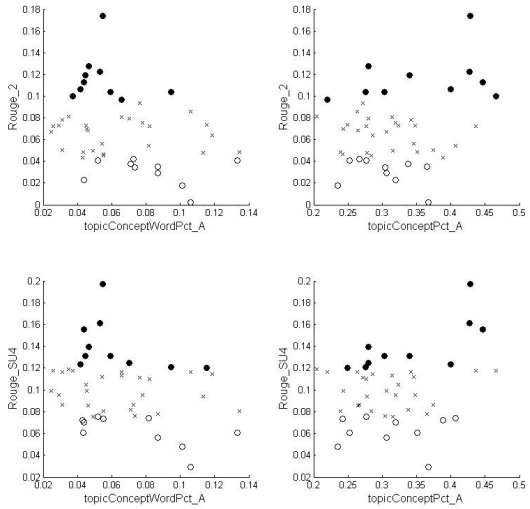

Fig. 4. System's topic ROUGE scores vs. topic's concept proportion. "topicConceptWordPct" means unique concept proportion in topic vocabulary, and "topicConceptPct" stands for total concept occurrence in topic length. • are the 10 best performed topics, ○ are the 10 worst, others are marked with ×.

the listed top concepts, we find that concepts in higher ranked topics are more correctly weighted, while the top concepts in lower ranked topics are more meaningless. For example, in D1024E-A, "Laden, United States, Sudan, National Isalamic Front" are the main characters involved in the bomb attack, and "Khartoum" is the location of the bomb event. But in D1023E-A, the central concept "Avalanche" is missing from the top list.

Looking into the factors which are responsible for concept weighting, we found the problem is due to an irrelevant context which is selected during wikification. For D1023E-A, totally 2618 keyphrases are found, in which 514 are unambiguous context concepts. The 26 context concepts with best context quality evaluated using Eq. 4 are: "Madeleine Albright, 23-Feb, Alan Greenspan, William Safire, Kosovo War, 10-Sep, 24-Feb, 8-Mar, Chuck Schumer, Tim Duncan, New York City, 12-Mar, 14-Jun, 30-Jun, Comprehensive Nuclear-Test-Ban Treaty, 15-Mar, Hillary Rodham Clinton, Kevin Garnett, Dennis Hastert, The New York Times, Brian Setzer, The Cincinnati Enquirer, Lauryn Hill, Vin Baker, Katharine Q. Seelye, Mohammad Khatami". Clearly, this context, which is full of names and dates, is not a good context for an article talking about avalanche. The reason for having so many names of Olympic athletes in this topic is because there are some "top news" documents in set D1023E-A. These documents contain multi-news about different events, one of which is avalanche.

For other poorly performed topics, the reason of bad ROUGE results are similar to what we have found in D1023E-A. Thus, we need to consider how to use concept relation to form a topic(local) focused context, than the current method which merely relies on concept relatedness in Wikipedia(global).

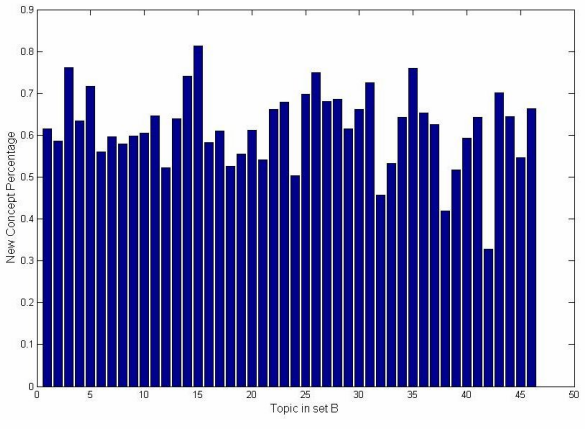

Fig. 5. New concept percentage in topics of set B

B. First paragraph information of Wikipedia

For set B, the main crucial factor is changed to feature WFP instead of WC, which actually are not used in the update task. This may due to WFP has already included the information in WC, which lead to a similar performance of set B with set A.

But WFP also provides more information using the relation to concepts' first paragraph, as specified in Eq. 12. Because of limited space, we list out only the top weighted concept and its Wikipedia article's first paragraph[1] for some topics, as shown in Table. V. For the topics, which have similar focus with the related concepts' first paragraphs, the performance shows a significant improvement. For example, topic D1027E-B is talking about a crime in which a homosexual boy Shepard was killed. The reason for his death was believed to be related to his sexual orientation. Thus this case caused a wide discussion and the adding of hate crimes laws. The top concept "Matthew Shepard" is correctly detected, which is actually just the topic title. The first paragraph of "Matthew Shepard" presents a detailed summary of this case, which can be even directly used for the topic(despite of its length). Topic D1013-B is in the similar case. Both of these two topics get the highest performance among all peers in ROUGE-2.

C. Detected new concept

Fig. 5 shows the new concept percentage in topics of set B. For example, D1015C-A talks about actions to protect rain forests taken by Africa, China and Malaysia. D1015C-B introduces actions in other countries, such as America, Brazil, and also an international effort. Concepts like "Forest", "Rainforest" and "Rain" are key concepts in both sets. While new concepts, e.g. "Amazon Rainforest", "Global warming", "Greenhouse effect", etc. show in set B.

But as feature WNC corresponding to a small weight in the updating summarization, it's not suitable to relate its

[1]We use the English Wikipedia from 9th Feb 2007 provided by JWPL for article content extraction.

TABLE I
TOPIC DISTRIBUTION OVER CATEGORIES

Id	Category Name	Topic percentage
1	Accidents and Natural Disasters	15.22%
2	Attacks (Criminal/Terrorist)	15.22%
3	Health and Safety	26.09%
4	Endangered Resources	21.74%
5	Investigations and Trials (Criminal/Legal/Other)	21.74%

TABLE IV
CONCEPTS IN EXAMPLE TOPICS OF SET A

Rank	TopicID	Title	Description	Concepts with top 10 WC weight	Concepts with top 10 context relation
1	D1024E-A	Bomb Khartoum	American attacked a factory in Khartoum saying it's associated with Laden. But Sudanese denied.	Osama bin Laden, Sudan, United States, United Nations, Muammar al-Gaddafi, Omar al-Bashir, Khartoum, National Islamic Front, Iraq, Saudi Arabia	National Islamic Front, Hassan al-Turabi, Omar al-Bashir, Osama bin Laden, John Garang, Taliban, Muammar al-Gaddafi, Arab League, Graf, Kofi Annan
2	D1043H-A	Rafik Hariri	Lebanese Prime Minister was killed by a group in Syria and Lebanon. France and U.S. meddled.	Rafik Hariri, Syria, Mosque, Saudi Arabia, Beirut, Politics of Lebanon, Bashar al-Assad, mile Lahoud, Jacques Chirac, Sunni Islam	Walid Jumblatt, Rafik Hariri, mile Lahoud, Walid Eido, Bashar al-Assad, Suleiman Frangieh, Parliament of Lebanon, Marwan Hamadeh, Arab League, Arab world
3	D1045H-A	Songhua River	Reactions of an explosion which caused pollution in Songhua River.	Songhua River, Benzene, Amur River, Northeast China, Heilongjiang, Ministry of Environmental Protection of the People's Republic of China, Pollution, Jilin, Water pollution, Harbin	Ministry of Environmental Protection of the People's Republic of China, Zhang Zuoji, Songhua River, Northeast China, Amur River, Benzene, China National Petroleum Corporation, Wen Jiabao, Nitrobenzene, Air pollution
4	D1033F-A	South Korean Wire Tapping	Investigation about illegally tapped conversation between Hong Seok-Hyun and Samsung official Lee Hak-soo.	Samsung Group, Roh Moo-hyun, Kim Dae-jung, JoongAng Ilbo, Hong Seok-hyun, Grand National Party, Democratic Party (Republic of Korea, 2005), South Korea, Telephone tapping, Yonhap	Democratic Party (Republic of Korea, 2005), Uri Party, Grand National Party, Hong Seok-hyun, Kim Seung-kew, Lee Hoi-chang, Kim Dae-jung, JoongAng Ilbo, Samsung Group, Lee Sang-ho
18	D1037G-A	Maryland Oysters	Dangerous situation of Maryland Oysters and protection steps of Maryland and Viriginia.	Oyster, Chesapeake Bay, Maryland, Tilghman Island, Maryland, United States Army Corps of Engineers, Algae, Virginia, United States, Grasonville, Maryland, Choptank River	Chester River, Choptank River, Tilghman Island, Maryland, Grasonville, Maryland, Chesapeake Bay, Rebecca T. Ruark (skipjack), Chesapeake Bay Bridge, Kent Narrows, J. Millard Tawes, Northern snakehead
30	D1039G-A	Murder Van Gogh	The reports and reactions to Netherland's first Islamic terrorist attack, in which Dutch filmmaker Van Gogh was killed.	Jan Peter Balkenende, Ayaan Hirsi Ali, Vincent van Gogh, Netherlands, Pim Fortuyn ,Amsterdam, Job Cohen, Rita Verdonk, Islam, Johan Remkes	Atzo Nicolai, Johan Remkes, Rita Verdonk, Jan Peter Balkenende, Piet Hein Donner, Pim Fortuyn, Ben Bot, Job Cohen, Ayaan Hirsi Ali, The Hague
33	D1030F-A	Ephedra	The benifits and risk of using Ephedra and restrictions in law.	Ephedrine, Ephedra, Herbalism, Metabolife, Caffeine, U.S. Food and Drug Administration, Cholesterol, Food, Stimulant, Amphetamine	Antidepressant, Methamphetamine, Amphetamine, Ephedrine, Caffeine, Major depressive disorder, Epinephrine, Insomnia, Ephedra, Asthma
44	D1014C-A	Obesity	Current situation of obesity in the world.	Obesity, Diabetes mellitus, Food, January 23, September 10, August 14, November 14, Hypertension, July 15, July 23	January 23, August 14, February 17, July 15, October 18, September 10, November 2, July 23, November 14, August 15
45	D1026E-A	Head Injuries	Helmets prevent many kinds of head injuries.	Myocarditis, Marfan syndrome, Anakinra, Overtraining, George Pataki, Automated external defibrillator, Commotio cordis, Hyperthermia, Hypertrophic cardiomyopathy, United States	Hypertrophic cardiomyopathy, Automated external defibrillator, Commotio cordis, Myocarditis, Long QT syndrome, Marfan syndrome, Hyperthermia, Anakinra, Gastroesophageal reflux disease, Skin cancer
46	D1023E-A	Austrian Avalanches	Damages of the avalanches in Austria and government's reactions.	United States, The New York Times, Bill Clinton, Austria, Kosovo, Lauryn Hill, February 23, European Union, National Basketball Association, Tim Duncan	Kevin Garnett, Tim Duncan, February 23, March 8, March 12, Vin Baker, March 15, June 30, June 14, Tim Hardaway

TABLE V
CONCEPTS IN EXAMPLE TOPICS OF SET B

Rank	TopicID	Title	Description
1	D1027E-B	Shepard Beating Death Trial	Trail of Shepard's murder, who was gay.
	[Top concept] Matthew Shepard [First paragraph] "Matthew Wayne Shepard" (December 1, 1976 - October 12, 1998) was an United States American student at the University of Wyoming, who was attacked by Russell Henderson and Aaron McKinney near Laramie, Wyoming, on the night of October 6 - October 7, 1998 in what was classed by many as a homophobic incident. Shepard died from his injuries at Poudre Valley Hospital in Fort Collins, Colorado, on October 12. Henderson is currently serving two consecutive Life imprisonment life sentences and McKinney is serving the same but without the possibility of parole. Henderson and McKinney were not charged with a hate crime - laws at the time did not support such a charge. Many people think the case should have been dealt with as a hate crime, particularly those who believe that Shepard was targeted on the basis of his sexual orientation. Under current federal United States law and Wyoming state law, crimes committed on the basis of sexual orientation are not prosecutable as hate crimes. Shortly after the murder, President Bill Clinton urged United States Congress Congress to add sexual orientation to the hate crimes law. The measure was defeated. In 1999, the Wyoming Legislature, amid widespread discussion of this crime, also attempted to pass legislation defining certain attacks motivated by victim identity as hate crimes, but the measure failed on a 30-30 tie in the Wyoming House of Representatives.	
2	D1044H-B	Red Food Dye	KFC and McDonalds stopped selling foods contains Sudan I sauce in China and Britain.
	[Top concept] KFC [First paragraph] "KFC", or "Kentucky Fried Chicken", is a fast food restaurant chain based in Louisville, Kentucky, United States. Founded by Colonel Sanders Colonel Harland Sanders and now a division of Yum!Brands, Inc., KFC is known mainly for its fried chicken. The company adopted the abbreviated form of its name in 1991 for three reasons: to deemphasize chicken (the Restaurant chain was moving to offer other foods), the unhealthy connotations of "fried", and the shorter name was considered more appealing to youth. Recently, the company has begun to remembrance the Kentucky Fried Chicken name, and now uses both "Kentucky Fried Chicken" and "KFC" in advertisements. The Kentucky Fried Chicken name can be seen on some buckets of chicken. As of 2006 KFC.com uses Kentucky Fried Chicken for the logo in the U.S.	
3	D1032F-B	Offshore Gas Leak	Statoil resumes production after the oil leak incident, which is due to careless.
	[Top concept] Statoil [First paragraph] "Statoil" (,) is a Norway Norwegian petroleum company established in 1972.It is the largest petroleum company in the Nordic countries and Norway's largest company. While Statoil is listed on both the Oslo Stock Exchange and the New York Stock Exchange, the Norwegian state still holds majority ownership, with 70.9%.The main office is located in Norway's oil capital Stavanger. The name Statoil is a truncated form of "the State's oil". Statoil is one of the largest net sellers of crude oil in the world, and a major supplier of natural gas to the European continent, Statoil also operates around 2000 gas station service stations in 9 countries. The company's CEO from mid-2004 onwards is Helge Lund, formerly CEO of Aker Kv?rner.In December 18th 2006 Statoil revealed a proposal to merge with the oil and gas division of Norsk Hydro, a Norwegian conglomerate. If the merger goes through, the new company will be the biggest offshore oil company in the world.	
16	D1013C-B	Identity Theft	US citizens are vulnerable to identity theft. Government is seeking for solutions.
	[Top concept] Identity theft [First paragraph] "Identity theft" is a term first appearing in U.S. literature in the 1990s, leading to the drafting of the Identity Theft and Assumption Deterrence Act. In 1998, The Federal Trade Commission appeared before the Subcommittee on Technology, Terrorism and Government Information of the Committee of the Judiciary, United States Senate. The FTC highlited the concerns of consumers for financial crimes exploiting their credit worthiness to commit loan fraud, mortgage fraud, lines-of-credit fraud, credit card fraud, commodities and services frauds. With the rising awareness of consumers to an international problem, in particular through a proliferation of web sites and the media, the term "identity theft" has since morphed to encompass a much broader range of identification-based crimes. The more traditional crimes range from dead beat dads avoiding their financial obligations, to providing the police with stolen or forged documents thereby avoiding detection, money laundering, trafficking in human beings, stock market manipulation and even to terrorism. The term "identity theft" is an oxymoron. It is not possible to steal a human identity. Human identity is a psychological thing, or the subject of philosophical conversation...	
44	D1026E-B	Head Injuries	Helmets prevent many kinds of head injuries.
	[Top concept] Motorcycle [First paragraph] "Motorcycle sport" is a broad field that encompasses all sport sporting aspects of motorcycle motorcycling. The disciplines are not all "races" or timed speed events, as several disciplines test a competitor's riding skills. The Fdration Internationale de Motocyclisme FIM is the international sanctioning body for motorcycle racing and most nations have their own governing bodies. Other breakaway/independent bodies exist on both a national and an international level. Motorcycles vary greatly in design for each different discipline. Broadly speaking, motorcycle racing can be divided into road racing or off road racing. Within each discipline there are usually sub-groups, including "Classic" machinery (i.e. non current to various extents), miniature machinery (e.g. Pocketbike Racing), "Sidecars" and "Quads / ATVs".	
45	D1001A-B	Columbine Massacre	Aftermath and blames of columbina massacre
	[Top concept] Eric Harris and Dylan Klebold [First paragraph] "Eric David Harris" (April 9, 1981; April 20, 1999) and "Dylan Bennet Klebold" (September 11, 1981 ; April 20, 1999), both high school seniors, were the perpetrators of the Columbine High School massacre in Jefferson County, Colorado Jefferson County, Colorado (near Denver, Colorado Denver and Littleton, Colorado Littleton), United States U.S., on Tuesday, April 20, 1999, killing 12 classmates and one teacher. Both Harris, 18 years old, and Klebold, 17, committed suicide after the killings.==Early life== Eric David Harris was born in Wichita, Kansas. His parents were Wayne and Kathy, and he had an older brother, Kevin. The family relocated often as Wayne Harris was a United States Air Force U.S. Air Force transport pilot; Kathy was a homemaker. ...	
46	D1009B-B	Jon Benét Ramsey	Investigations going on for the murder.
	[Top concept] JonBenét Ramsey [First paragraph] "JonBenét Patricia Ramsey" (August 6, 1990; December 25, 1996) was a beauty contest beauty pageant contestant who was found murdered, at the age of 6, in the basement of her parents' home in Boulder, Colorado, nearly eight hours after being reported missing. The case drew national attention in the United States when no suspect was charged and suspicions turned to possible family involvement. The tantalizing clues of the case inspired numerous books and articles that attempt to solve the mystery. On August 16, 2006, the case returned to the news when John Mark Karr, a 41-year-old school teacher, reportedly confessed to her murder. On August 28, 2006, the district attorney, Mary Keenan Lacy, announced that Karr's DNA did not match that found at the scene, and no charges would be brought against him.	

performance with the ROUGE score. Thus we leave the analysis of this feature for future work.

VII. Conclusion

In this paper, we present our on-going work on multi-document summarization using concepts and their articles' first paragraph in Wikipedia. From the analysis of our system's performance in TAC 2010 evaluation, we are optimistic of the potential to use Wikipedia for the summarization task.

Intuitively, meaningful and unambiguous concepts are much more useful for text analysis than ambiguous and noisy words. In our experiments, we found that the number of detected concepts doesn't significantly affect system performance. This allows the system to be more robust. For all topics in TAC2010 dataset, a sufficient number of Wikipedia concepts has been detected, especially the most important one. Apart from concepts, the first paragraph in Wikipedia articles proved to be especially suitable for the summarization task. As shown in experiment, the system's performance is highly improved with the support of a closely related paragraph.

In order to fully mine Wikipedia's potential for summarization, there are some problems that need to be solved:

- Context representation. Context is the base of concept disambiguation and concept selection. A misleading context will cause the system to choose wrong senses of ambiguous concepts and blur the focus of resulting summary.
- Category information. Wikipedia provides a huge category structure for concepts. We believe the usage of concept category will be helpful when dealing with the multi-subtopic problem in multi-document summarization. Categories are also useful for concept relatedness assessment, which is elementary for concept analysis.
- Supporting information for a summary. The quality of first paragraph varies for each topic, which leads to an unstable system performance. Further exploration through Wikipedia is needed to find other kinds of supporting information for summarization.

We leave these problems for our future work.

Acknowledgment

This work is partially supported by National Natural Science Foundation of China (10871019/a0107) and the Key Project of Chinese Ministry of Education(108126).

References

[1] D. Milne and I. H. Witten, "Learning to link with wikipedia," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 509–518.

[2] R. Mihalcea and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York, NY, USA: ACM, 2007, pp. 233–242.

[3] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang, "MEAD - a platform for multidocument multilingual text summarization," in *LREC 2004*, Lisbon, Portugal, May 2004.

[4] C.-Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Proc. of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain*, 2004.

[5] E. Hovy, C. yew Lin, L. Zhou, and J. Fukumoto, "Automated summarization evaluation with basic elements," in *In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC*, 2006.

[6] R. P. Ani Nenkova, "Evaluating content selection in summarization : The pyramid method," in *HL T2NAACL 2004*, 2004.