RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D5.2.1

Definition of evaluation metrics and first prototype of diversified news service

Editor:	Enrique Alfonseca, Google Inc.
Author(s):	Enrique Alfonseca, Google Inc., Jean-Yves Delort, Google Inc.
Deliverable Nature:	Report (R)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	September 2011
Actual Delivery Date:	September 2011
Suggested Readers:	Developers involved in the use cases.
Version:	2.1
Keywords:	summarization

Disclaimer

This document contains material, which is the copyright of certain RENDER consortium parties, and may not be reproduced or copied without permission.

In case of Public (PU):
All RENDER consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):
All RENDER consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):
The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement. However, all RENDER consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):
The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the RENDER consortium as a whole, nor a certain party of the RENDER consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	RENDER – Reflecting Knowledge Diversity
Short Project Title:	RENDER
Number and Title of Work package:	WP5 Diversity case studies
Document Title:	D5.2.1 - Definition of evaluation metrics and first prototype of diversified news service
Editor (Name, Affiliation)	Enrique Alfonseca, Google Inc.
Work package Leader (Name, affiliation)	David Cadenas Sanchez, Telefonica

Copyright notice

Executive Summary

The Google use case is mainly focused on application of diversity-aware text summarization to use cases that are relevant to the company.

The deliverable includes three main sections: a description of the main use case, consisting of a diversity-aware news summarization system; a technical description of the particular technical solutions chosen for the summarizer and implemented so far, including some examples of generated summaries; and the way we intend to evaluate the systems.

One of the summarization systems developed internally has been used in a public competitive evaluation at the TAC-2011 competition.

Table of Contents

List of Figures and/or List of Tables

Abbreviations

KL	Kullback-Leibler
LDA	Latent Dirichlet Allocation
TAC	Text Analysis Conference

1 Introduction

Google collects articles from about 25,000 professional sources for Google News, and additionally also serves items provided by non-professional Web users such as blog items or tweets. This information is made available to our users through our products, including Web Search, Google News and Google Alerts. In all of these products, users specify their information needs or preferences by means of a short textual query, that we use to choose the most relevant documents to be returned.

The focus of the RENDER project is the analysis and representation of diversity of opinions in the web. News articles is one source of diversity, as the same event can be described in news sources from different viewpoints. We believe that users would benefit from having these viewpoints explicitly presented to them, making it easier for them to analyse the underlying events and to form an informed opinion about them. To this aim, a variety of existing technologies can be combined and further developed so that news textual sources can be analyzed and contrasted with each other, the writer’s attitude (and polarity) toward the event can be coarsely classified, and the most relevant facts and viewpoints can be distilled from the collection of articles and presented to the user in a useful way so that the reader can make an informed, critical judgement of the event described.

This document starts by describing the Google use-case specification. This is followed by a description of the currently working prototype and how we plan to extend and improve it going forward. Finally, we describe the evaluation settings and metrics we are considering to measure the success of this research.

2 Use-case description

For the use case with news, we plan to develop a diversity-aware visualization tool for the relevant information returned to the user. For example, for a user looking for [libya] in Google News, the current result is a set of news, those referring to the same story grouped together, about the general topic of the country Libya. This visualization tool should be able to process the news that refer to the same events, and to show to the users additional ways for analyzing the information, as for example:

- What is the most relevant information about this entity for the use case of the user? For example, if a user is searching for a recent event, we would produce a summary of the most recent news, whereas a user searching for an entity for which there is not a very important recent interest, we could show important information about it from a time span going farther back into the past. This first problem is similar to the general query-guided summarization [Dang, 2006], where a user specifies a query containing an information need, and summarizer outputs are evaluated on several dimensions, including how informative they are to the user and how readable, coherent and grammatical they are.

- What are the different points of view in the different news that refer to the same event at different times? Different sources may report the same data with differences in bias or sentiment, or may simply highlight and omit different characteristics of the news. There is already substantial work combining sentiment analysis [Pang and Lee, 2008] and text summarization [Mani, 1999]. Some early approaches combining summarization with sentiment include [Stoyanov et al., 2004] and [Yu and Hatzivassiloglou, 2003], which divided the task into two subtasks: first finding the most relevant sentences in the corpus given the user query, and next classifying them as either factual or opinionated, in order to keep only the opinionated ones. The Text Analysis Conference in 2008 included a competitive evaluation on opinion mining, where most of the systems approached the problem by adapting their summarizers to include in sentence scoring a polarity or subjectivity score, or by filtering out non-subjective sentences. A related area is that of aspect-based summarization, which tries to learn automatically, given a set of user reviews, what are the most relevant aspects mentioned in the review that users care about, the polarity of user opinions about each of them, and short snippets of text. [Titov and McDonald, 2008; Lu et al., 2009; Zhu et al., 2009].

- How are these news related to news that happened in previous time intervals? Is it a new development on an already known event, or is it an entirely new event? What are the data contained in the news that is general background knowledge, information that was presented in the past, and new information presented now? All these are specific research questions that can be used to improve the presentation of the summary to the user, and provide power users additional flexibility for tuning the summaries to their specific needs.

Summaries can be presented in many different ways. The most traditional form of summaries consist of textual, natural language content. However, for some applications, using short snippets (which may not include complete, grammatical sentences), or graphics, may be a more efficient way of representing and compacting the information. We consider the visualization itself as a research question, and solving it will necessarily involve user experience analyses and trials with real users. Timelines, tables, graphics and textual summaries are means that we are initially considering as possibly useful for this task.

The techniques needed to implement this showcase constitute research problems that have been partially explored by the research community working on areas such as Sentiment Analysis, Opinion Mining, Text Summarization or Topic Detection and Tracking. It is our purpose to make progress in the state-of-the-art in some of these fields while evaluating how useful these techniques are in a real use case scenario.

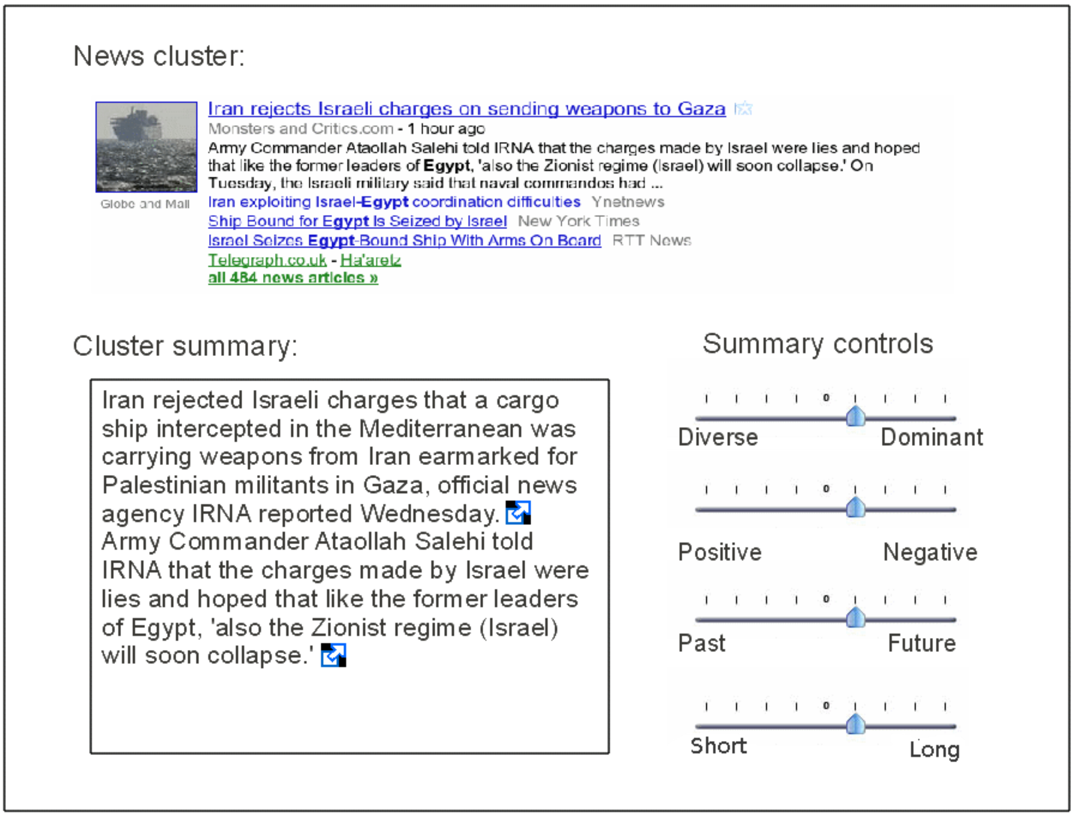
Figure 1 displays a screenshot of a possible user interface:

- A news cluster, containing several news articles describing the same event, needs to have been previously selected by the user. The system will work with the news included in the cluster, which should all be talking about the same or very related events.

- In this mock-up, a text-based summary is generated from the different documents in the cluster. Different ways of visualization would include sentiment analysis color-based visualization (indicating the proportion of news writers that have a positive or a negative stance towards some event) or timelines.

- Some controls will allow the user to tune the summary. Regarding diversity, two very relevant ones will be (a) the ability to be able to see the most dominant interpretation of the event, or to see a wider variety of opinions reflected in the summary, and (b) the ability to generate more positive or more negative summaries, depending on the polarity of the authors. As indicated, we could also tune the summary towards the presentation of past events or speculation about the future. Finally, a common feature in summary generation systems is the possibility of generating summaries of different lengths.

In a second iteration, we plan to extend the previous summaries with temporal information: instead of summarizing a single event (represented by a set of news) we plan to put the event in context with respect to other previous related events, under settings similar to topic detection and topic tracking.



News cluster:

Iran rejects Israeli charges on sending weapons to Gaza
Monsters and Critics.com - 1 hour ago
Army Commander Ataollah Salehi told IRNA that the charges made by Israel were lies and hoped that like the former leaders of **Egypt**, 'also the Zionist regime (Israel) will soon collapse.' On Tuesday, the Israeli military said that naval commandos had ...
Iran exploiting Israel-**Egypt** coordination difficulties Ynetnews
Ship Bound for **Egypt** Is Seized by Israel New York Times
Israel Seizes **Egypt**-Bound Ship With Arms On Board RTT News
Telegraph.co.uk - Ha'aretz
all 484 news articles »

Cluster summary:

Iran rejected Israeli charges that a cargo ship intercepted in the Mediterranean was carrying weapons from Iran earmarked for Palestinian militants in Gaza, official news agency IRNA reported Wednesday. Army Commander Ataollah Salehi told IRNA that the charges made by Israel were lies and hoped that like the former leaders of Egypt, 'also the Zionist regime (Israel) will soon collapse.'

Figure 1: Mock-up of the summarization user interface.

3 Summarization system

We have implemented a prototype of the summarization system which is able to generate generic summaries of clusters of related news articles. The summaries generated are all textual, written in natural language. This section contains a more detailed description of the techniques used to produce them. Some examples of the output summaries automatically generated by this first prototype are provided in Figures 2, 3 and 4.

The different summarization models implemented are the following:

- **NistSum:** this is a very simple summarizer, commonly used as a baseline in the summarization competitions organized by NIST, which simply selects the beginning of the most recent document in the cluster of news articles. Usually, news articles starts by stating the most important information about the event and next proceed by providing more details and some background. Therefore, when the quality of the news article is good, this simple heuristic usually generates a good summary. Some of the drawbacks is that it is very dependent on the quality of the last article in the cluster, the results are not stable with respect to small changes in the collection (in particular, if the most recent news changes in the collection the summary is completely different), and it does not capture different points of view or different sub-events that may be stated in different articles.

- **SumBasic (Nenkova and Vanderwende, 2005):** this is a simple summarizer based on the observation that the most common words in a document collection (excluding stopwords) are more likely to appear in human-generated summaries. It proceeds in the following way: first of all, a unigram language model $P_D(\cdot)$ is obtained from the source document collection D, e.g. using the maximum likelihood estimate. This language model will assign a probability to each word in the vocabulary between 0 and 1, such that the sum of all probabilities is 1:

  $$\sum_{w \epsilon V} P_D(w) = 1$$

  Every sentence S is then scored as the average of the probabilities assigned to all the words w that appear in that sentence:

  $$Score(S) = \sum_{w \epsilon S} \frac{1}{|S|} P_D(w)$$

  In a third step the sentence with the maximum score is chosen to appear in the summary, and the language model $P_D(\cdot)$ is updated by squaring the probabilities of all the words that appeared in the selected sentence. This effectively reduces the probability of choosing a very similar sentence later on. Finally, all sentences are re-scored with the updated language model, and the process continues. When the length limit is reached, the algorithm stops. In practice, despite the squaring of the selected content words, it is not very good at removing redundancy.

- **KLSum (Haghighi and Vanderwende, 2009):** given the language model of the document collection $P_D(\cdot)$, every possible extracted summary is evaluated as

  $$Score(Sum) = KL(P_D \| P_S)$$

where KL is the Kullback-Leibler divergence, and P_S is the empirical unigram distribution from the summary. The summary that minimizes the score is the one whose unigram distribution is closest to the collection's, and will be chosen as the system output. Because the space of possible summaries is combinatorial in the total number of sentences in the collection, evaluating all of them is intractable in practice and a greedy approach is commonly followed.

- **TopicSum (Haghighi and Vanderwende, 2009):** TopicSum uses a simple LDA-like topic model (Blei et al., 2003) to better estimate the collection topic model used in KLSum.

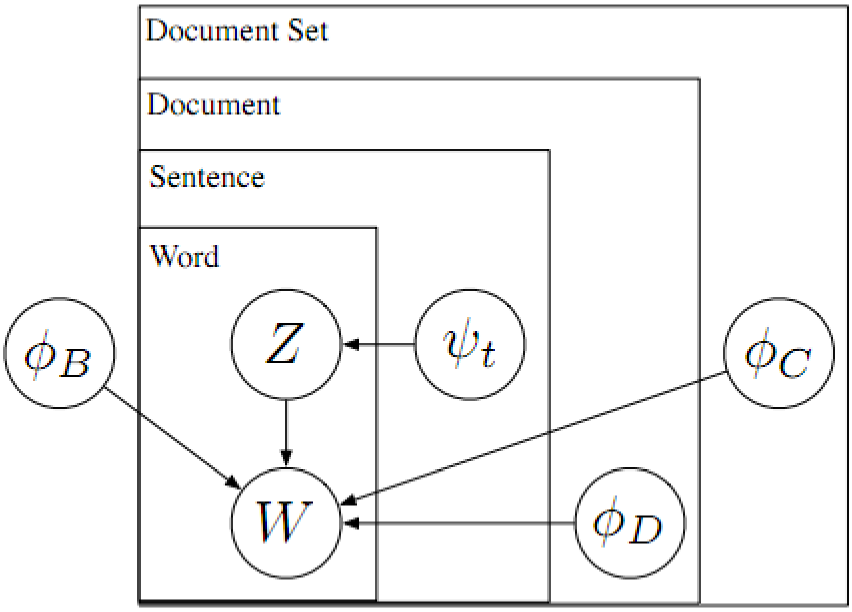


Figure 2: TopicSum plate diagram

Figure 2 shows the plate diagram. The objective is to be able to identify which words belong to (a) a background distribution ϕ_B containing general words that can appear in any document about any topic, such as stopwords and auxiliary verbs; (b) per-document distributions ϕ_D containing the most characteristics words in that document; and (c) per-collection distributions ϕ_C containing the most characteristics words in each collection. Given a particular document collection C to be summarized, its distribution ϕ_C will be used as the collection distribution in KLSum. In the diagram, W refers to a word, and Z is the selector variable that decides whether the word is chosen from the document distribution, the collection distribution or the background distribution. The mixture of topics inside a document that determines the topic probabilities that govern Z is ψ_t.

- **DualSum:** we have developed a new topic model as a modification of TopicSum for the particular case of update summaries: when the user already knows about some previous facts regarding an ongoing event, and our task is to generate a summary highlighting the novelty of the new information received about the unfolding event. For example, if a user is interested about the the drought in Somalia and has read news about it until two days ago, it is possible to generate a summary from the collection of news from the last two days. However, these late news are likely to contain background information that the user already knows, and it should not be necessary to add it to the summary. The plate diagram in Figure 3 shows a topic model built specifically for this purpose: each collection is divided into two parts: the old information about the event, and the new articles that were last received. The words in these new articles are allowed to be generated from the new collection language model or from the old collection language model for the event, thus allowing recent articles to mention background information. The notation is the same as in TopicSum: w refers to a word, z to a topic assignment for that word, and words can be selected using either the background distribution ϕ_B, the collection distributions ϕ_C and $\phi_{C'}$, and the document distributions ϕ_D and $\phi_{D'}$. Furthermore, the words in the updated (new) collection can also be chosen from the collection distribution from the original (old) documents, ϕ_C.

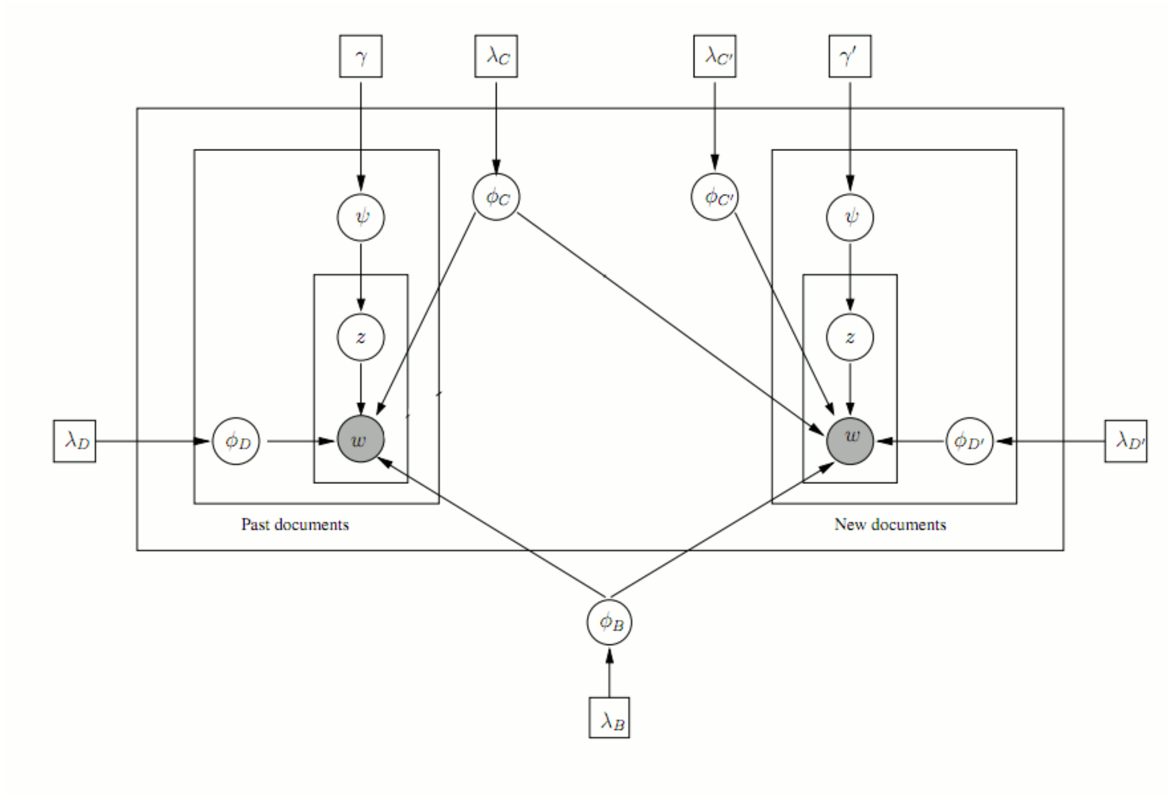


Figure 3: DualSum plate diagram

Using this new system we have also participated in the update summarization track in the TAC-2011 competitive evaluation (TAC, 2011). This is a competitive evaluation where different systems participate applying their system on the same set of document collections, and manual and automatic evaluations are done on the resulting outputs, to be able to identify the best technologies for the task. In 2011 there were the following tasks: (a) **guided summarization**, consisting in generated a query-focused summary of a collection, finding the information in the summaries that is most important given the user's query; (b) **guided summarization, update task**, consisting in generating a query-focused summary of a collection, finding the information in the summaries that is most important given the user's query, and that is not redundant given a set of old news that are known about the same topic; (c) **summary evaluation**, to explore the best strategies in automatically evaluating the output of summarization systems; and (d) **multi-lingual evaluation**, aiming at developing algorithms and resources that are applicable across languages. Our participation in the competition was in the update task. Results are not released yet.

Figures 4, 5 and 6 show example summaries generated with these systems for real clusters of news articles, together with a title and snippet for a selection of the articles. These have been generated using TopicSum, by extracting the sentences that are most similar to the collection language models.

HP resurrects TouchPad tablet to pacify rabid customers

By Mike Isaac, WIRED HP has plans to produce another round of its TouchPad tablets before the year is out, despite its earlier decision to discontinue its mobile hardware products. "Despite announcing an end to manufacturing webOS hardware, **...**

Last ever batch of TouchPads isn't coming to Blighty

By Bill Ray • Get more from this author Brits hoping for a cheap fondle won't get one from HP, as the final production run of the TouchPad will be bound for North America only. We knew one more production run was planned, but mobot.net managed to get **...**

Ahead of the Bell: HP restarts tablet production

AP , 08.31.11, 08:52 AM EDT NEW YORK -- Hewlett-Packard Co.'s decision to restart production of its TouchPad tablet is "confusing," a Sterne Agee analyst said Wednesday, but it could improve the value of the webOS software used by the tablet, **...**

HP resurrects TouchPad for one last go at the iPad

Jon Rubinstein, senior vice president and general manager for Palm, holds the Palm TouchPad during a media presentation at the Herbst Pavilion at the Fort Mason Center in San Francisco February 9, 2011. LOS ANGELES (Reuters) - Hewlett Packard Co (HPQ.**...**

HP update to boost TouchPad 'functionality'

Hewlett-Packard told CNET on Monday that it will deliver a future update to boost the TouchPad's functionality. HP also said it has seen "huge spikes" in TouchPad activation in the wake of the decision to discount the tablet. **...**

HPs TouchPad: The thing that would not die!

Now, in a case of strategic whiplash, an HP executive says the TouchPad could be reborn if another part of HP's restructuring plan comes to pass. Todd Bradley, who heads up HP's Personal Systems Group, told Reuters that if HP does indeed spin off its **...**

Modders Slap Popular Android Hack On HP's TouchPad

By Mike Isaac It was only a matter of time before the hacks for HP's now defunct tablet started to roll in. Android-modding group CyanogenMod released a video of its popular aftermarket software running on HP's TouchPad tablet, a product which normally **…**

Summary

HP plans another production run of TouchPad tablets to meet the demand for the $99 product. It is to produce just one last batch of its aborted TouchPad tablet line, the company said on Tuesday. Just two weeks before, retailers wanted to give HP back all the TouchPads they had in stock.

Figure 4: Cluster of news with snippets, and summary generated from it.

Poor sleep increases risk of high blood pressure

New research reports that poor sleep quality raises the risk of elevated blood pressure in elderly men. According to the study, published Aug. 29 in Hypertension: Journal of the American Heart Association, lack of deep sleep, or slow wave sleep (SWS), **...**

Lack of Deep Sleep Tied to Hypertension

By ANAHAD O'CONNOR Men who get the least deep sleep each night have a higher risk of hypertension, new research shows. Earlier studies have tied chronic sleep disorders and low levels of sleep to greater risks of heart disease and obesity, **...**

Poor sleep increases high blood pressure risk

By Kathleen Doheny, HealthDay If you sleep poorly, your chances of developing high blood pressure may increase, new research suggest. Elderly men who spend little time in deep sleep could be at risk of developing high blood pressure. **...**

Could Lack of Deep Sleep Fuel High Blood Pressure?

By Anne Harding/Health.com Tuesday, August 30, 2011 | View Comments Missing out on deep sleep can leave you feeling slow-witted and irritable in the morning, but the consequences don't necessarily end there. Over time, too little deep sleep may also **...**

Bad sleep ups blood pressure risk

Elderly men who spend little time in deep sleep could be at risk of developing high blood pressure, according to US scientists. A study on 784 patients, in the journal Hypertension, showed those getting the least deep sleep were at 83% greater risk **...**

Poor Sleep May Raise Blood Pressure

By Jennifer Warner A new study shows men who got the least deep sleep were 80% more likely to develop high blood pressure than those who got the most. Researchers determined how much deep sleep the men got by measuring the speed of their brain waves. **...**

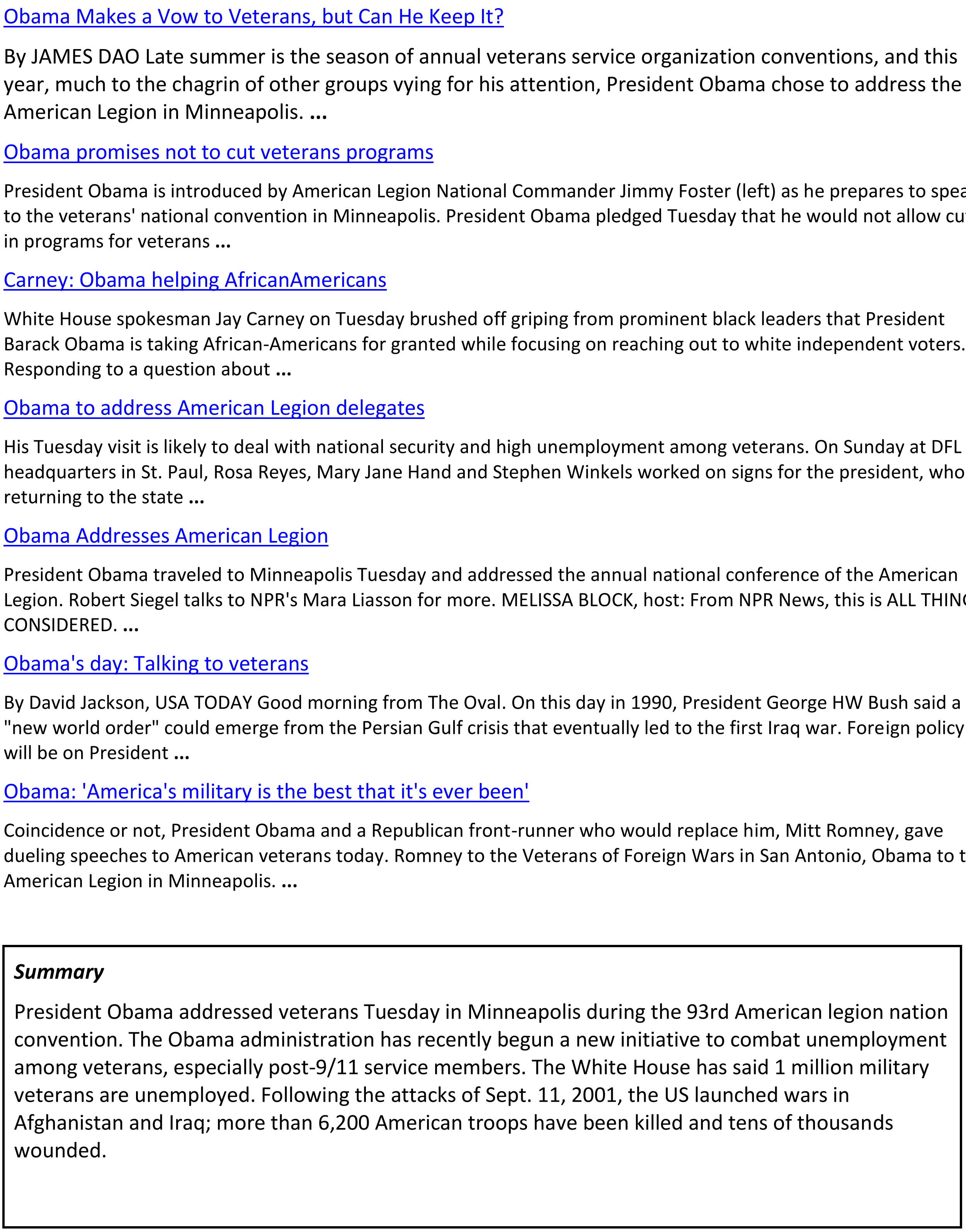
Deep Sleep Dreamers Enjoy Normal BP

By Nancy Walsh, Staff Writer, MedPage Today Note that slow wave sleep, a stage of non-REM sleep, is considered to be restorative and is the stage associated with the highest arousal threshold. Point out that in this study, older men who were **...**

Summary

The researchers said that older men who spend less sleep time in slow-wave sleep have an increased risk of developing high blood pressure. The study is published in Hypertension, a journal of American Heart Association. Doctors at Harvard Medical School looked at men around age 75. Good quality sleep is the third pillar of health, Redline said.

Figure 5: Cluster of related news, and summary generated from it.



Obama Makes a Vow to Veterans, but Can He Keep It?

By JAMES DAO Late summer is the season of annual veterans service organization conventions, and this year, much to the chagrin of other groups vying for his attention, President Obama chose to address the American Legion in Minneapolis. **...**

Obama promises not to cut veterans programs

President Obama is introduced by American Legion National Commander Jimmy Foster (left) as he prepares to speak to the veterans' national convention in Minneapolis. President Obama pledged Tuesday that he would not allow cuts in programs for veterans **...**

Carney: Obama helping AfricanAmericans

White House spokesman Jay Carney on Tuesday brushed off griping from prominent black leaders that President Barack Obama is taking African-Americans for granted while focusing on reaching out to white independent voters. Responding to a question about **...**

Obama to address American Legion delegates

His Tuesday visit is likely to deal with national security and high unemployment among veterans. On Sunday at DFL headquarters in St. Paul, Rosa Reyes, Mary Jane Hand and Stephen Winkels worked on signs for the president, who is returning to the state **...**

Obama Addresses American Legion

President Obama traveled to Minneapolis Tuesday and addressed the annual national conference of the American Legion. Robert Siegel talks to NPR's Mara Liasson for more. MELISSA BLOCK, host: From NPR News, this is ALL THINGS CONSIDERED. **...**

Obama's day: Talking to veterans

By David Jackson, USA TODAY Good morning from The Oval. On this day in 1990, President George HW Bush said a "new world order" could emerge from the Persian Gulf crisis that eventually led to the first Iraq war. Foreign policy will be on President **...**

Obama: 'America's military is the best that it's ever been'

Coincidence or not, President Obama and a Republican front-runner who would replace him, Mitt Romney, gave dueling speeches to American veterans today. Romney to the Veterans of Foreign Wars in San Antonio, Obama to the American Legion in Minneapolis. **...**

Summary

President Obama addressed veterans Tuesday in Minneapolis during the 93rd American legion nation convention. The Obama administration has recently begun a new initiative to combat unemployment among veterans, especially post-9/11 service members. The White House has said 1 million military veterans are unemployed. Following the attacks of Sept. 11, 2001, the US launched wars in Afghanistan and Iraq; more than 6,200 American troops have been killed and tens of thousands wounded.

Figure 6: Cluster of related news articles and summary generated from it.

4 Evaluation metrics

We are going to need two different kinds of evaluation metrics during the development of this work:

- During development, to allow for a fast turnaround and a short development cycle, we need an automated evaluation that can be quickly run. For this, the most common evaluation metric used in the summarization field is ROUGE-2, with other similar alternatives having been suggested as well. We have licensed the previous datasets from the Document Understanding Conference and the Text Analysis Conference summarization tracks, and we have built the infrastructure to evaluate using those datasets and the ROUGE-2 score. As the project progresses we may add or remove datasets and decide on a different evaluation metric.

  There are many other evaluation scores that have been proposed over time to evaluate automatic summarization systems, but none of them has been yet as widely accepted as ROUGE-2. Although ROUGE-2 has some well-known problems, mostly stemming from the fact that it is an n-gram recall score with respect to manual summaries, which makes it harder for ROUGE-2 to properly evaluate abstracts or tasks when there is not much lexical overlap between the generated summary and the manual reference summaries, so far there is no commonly agreed replacement for it as a fully automatic evaluation metric.

- At certain milestones we will evaluate the summaries with an application-specific evaluation. Having in mind the application of generating summaries for news clusters, an evaluation procedure that is very amenable is to compare different summarizers using side-by-side experiments. Under these settings, the human raters will be able to read the articles from a cluster of related news, and then they will be presented two alternative summaries, generated with different implementations of our summarizer. The task will be to choose which summary is more useful to them when browsing news. Having the evaluation defined as binary decision makes it easier for the raters to understand the task and hopefully we will be able to have substantial agreement between them.

  This same approach can be used to evaluate other applications. One application that we have in mind is Google Alerts. In this application, users sign up to receive an email every day or every week containing the latest news, blog posts or web results about a certain topic of interest. The user interest is expressed as a keyword-based query. The emails are generated automatically, and contain a list of links pointing to new documents that the system believes the user will be interested in. The plan here is to evaluate whether a textual summary would be considered more useful to the users than the list of links to pages.

  In side-by-side experiments, users are shown, at two sides of the screen, the output of two different systems. They are usually designed to compare the current version of a system to a possible enhancement, to make sure that the users find the new one better than the old one. The order of the two system outputs is randomized, so different users may see them in different order, and the different rating tasks shown to the same user may also be presented in a different order. This ensures that no ordering bias is present resulting in a preference for any system. The raters are chosen using a crowdsourcing mechanism, so they are not necessarily trained for the task and it is important to describe the evaluation guidelines very clearly, as otherwise a failure to understand the task may result in invalid results. It is also convenient to make sure that no single rater rates more than a fraction of the total evaluation set, say 3% or 5%, to guarantee that the personal view of that rater is not biasing the overall evaluation. A post-processing of the data can also be done to guarantee that there was a minimum inter-rater agreement and, if that minimum agreement is not reached, the evaluation guidelines or template has to be changed in a new rating iteration, to ensure that the raters really understood the rating task and that this task is objective and well defined.

5 Interaction across the consortium

One of the core problems to solve In the detection and conveyance of diversity in textual sources is how to present this information to the users. While it can be presented as overlayed annotations in the source documents, we believe that text summarization is a natural way to present this information straight away in natural language, where free text can be used to indicate which documents contain some bias of which the reader should be aware, and which are the points in which two or more news articles describing the same event differ.

DualSum, the algorithm proposed for update summarization (highlighting novel information in a situation in which the user has already read previous articles), can be used to identify and highlight the main differences between any two sets of news. One of the current applications that we are exploring is the use of DualSum to find the main differences between the main Wikipedia entries and their associated discussion pages, in order to discover which are the most controversial topics that are discussed by the editors of Wikipedia about each entry. We also plan to extend this work to automatically identify discussion-raising statements in other Wikipedia entries that may not have given rise to discussion yet. In this way we expect the research developed for Google's use case to be applicable to solve some of the problems presented for the Wikipedia use case.

At the same time, we are now closely collaborating with JSI in the identification of related news which describe the same event as it develops over time. This will allow us to get more familiarity with the tools developed by JSI in the context of the project and to jointly develop some of them. The expected result of this collaboration will be a method to identify which different news refer to the same event (as it unfolds over time), and to enhance them with a variety of linguistic annotations. A follow-up of this work will be the study of how to present this information to the user in the most effective way.

6 Conclusions and Future Work

This document describes the first prototype for news summarization, and the evaluation metrics for the generated summaries. The current prototype is able to generate different kinds of summaries using various techniques, including generic multi-document summarization (generating a summary from a set of news articles), and update summarization (generating a summary that highlights what is the new information mentioned assuming that the user already knows about a previous set of articles). There is an existing demo built that a user can use to create summaries of existing clusters of news. Currently this is internal for Google.

As future work, the first step will be to incorporate some controls to be able to tune the summary. Some of them include to adapt the summary to different polarity requirements, highlighting what are the most positive facts mentioned in the original articles and what is the negative information. A second enhancement will include the identification of diversity of opinions, by choosing opinionated sentences and identifying the most relevant ones. In parallel, we are exploring applications of the new technology to the Wikipedia use case.

Google is currently collaborating with Wikipedia and JSI in order to advance with these future plans.

References

Blei et al., 2003]
Blei, D. M., Ng, A. Y. and Jordan, M. I., *Latent dirichlet allocation*. JMLR (2003)

[Dang, 2006]
Trang Dang, H., *DUC 2005: Evaluation of question-focused summarization systems*. In Proceedings of the COLING-ACL'06Workshop on Task-Focused Summarization and Question Answering, pages 48–55, Prague, 2006.

[Haghighi and Vanderwende, 2009]
Haghighi, A. and Vanderwende, L., *Exploring content models for multi-document summarization.* In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009.

[Mani and Maybury, 1999]
Mani, I. and Maybury, M. T., *Advances in automatic text summarization*. The MIT Press, 1999.

[Nenkova and Vanderwende, 2005]
Nenkova, A. and Vanderwende, L., *The impact of frequency on summarization*. Technical report, Microsoft Research (2005).

[Pang and Lee, 2008]
Pang, B. and Lee, L., *Opinion Mining and Sentiment Analysis*. In Foundations and Trends in Information Retrieval, volume 2, issue 1-2., January 2008.

[Stoyanov et al., 2004]
Stoyanov, V., Cardie, C., Litman, D. and Wiebe, J., *Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus*. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (2004).

[TAC, 2011]
TAC 2011 Summarization Track. http://www.nist.gov/tac/2011/Summarization/ Proceedings to appear at the Text Analysis Conference (TAC) 2011 Workshop, November 14-15, 2011, National Institute of Standards and Technology, Gaithersburg, Maryland USA

[Titov and McDonald, 2008]
Titov, I. and McDonald, R., *A joint model of text and aspect ratings for sentiment summarization*. In Proceedings of ACL-2008.

[Yu and Hatzivassiloglou, 2003]
Yu, D. and Hatzivassiloglou, V., *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. In Proceedings of EMNLP-2003.

[Yue et al., 2009]
Yue, L., Zhai, C. and Sundaresane, N., *Rated aspect summarization of short comments.* In Proceedings of WWW-2009.

[Zhu et al., 2009]

Zhu, J., Zhu, M., Wang, H. and Tsou, B. K., *Aspect-based sentence segmentation for sentiment summarization.* In Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, 2009.