Cornell Expert Aided Query-focused Summarization (CEAQS): A Summarization Framework to PoliInformatics

Lu Wang*, Parvaz Mahdabi◇, Joonsuk Park*, Dinesh Puranam†, Bishan Yang*, Claire Cardie*
*Department of Computer Science, Cornell University
{luwang, jpark, bishan, cardie}@cs.cornell.edu
◇Faculty of Informatics, University of Lugano, Switzerland‘
parvaz.mahdabi@usi.ch
†The Samuel Curtis Johnson Graduate School of Management, Cornell University
dp457@cornell.edu

Abstract

We present a query-focused summarization framework to extract salient information for the PoliInformatics unshared task according to user-specified queries. Our system calculates sentence importance based on word frequency, speaker expertise, and topic relevance. Temporal changes in topic and speaker importance can be identified with our system summaries.

1 Introduction

The PoliInformatics dataset contains multiple types of text-based resources, such as Federal Open Market Committee (FOMC) and Financial Crisis Inquiry Commission (FCIC) meeting transcripts and U.S. Congressional hearings. Because each resource provides information on similar topics associated with the financial crisis of 2007–2008, we hypothesize that the resources can provide complementary information on each topic. Thus, our system aims to combine information from these sources for a more complete representation of each topic[1].

The system works as follows: First, the dataset is segmented into meta-documents without making distinction among source types. Then, a ranked list of snippets from relevant documents for each topic of interest are retrieved by an information retrieval system, making use of the queries generated for the topics. To improve the summarization results of the basic system (SUMBASIC) we model “expertise”. Using topic models we identify “expert” speakers for each topic and generate scores for each speaker. The basic system, augmented with these scores, retrieves snippets that weighs the expertise for each speaker while presenting results. Another method to model expertise is to use topic models to identify topic-relevant/expert sentences to represent each document/speech returned by the summarization system. The rest of this paper presents the details of the system and the results. Our system output is available at http://www.cs.cornell.edu/~luwang/polinformatics/main.html

2 Pre-processing

Segmentation: We first segment the dataset into individual text units. The granularity for the segmentation is chosen as follows: wherever the textual content is pre-processed by organizers and we have access to extracted speaker information in CSV files, we use this information and consider each utterance, i.e., text snippet, associated with a speaker as a separate text unit. This is the case for the data in folders such as: FOMC and Congressional hearings. However, there are resource types for which we do not have access to such CSV files. In such cases, we decided to identify speaker information using regular expressions that rely on capitalization. Please note that we did not use the congressional bills in our system.

Indexing: We indexed the text units (hereafter refer to as document) using Terrier[2]; performed stemming using the Porter stemmer; and removed stop-words according to Terrier’s default stopword list. Table 1 reports statistics on the index.

Number of Documents	41,204
Number of Tokens	3,520,096
Number of Unique Terms	24,828

Table 1: Statistics about the Index

Query Formulation: We used the Wikipedia page for the “Financial crisis of 2007-08”[3] to get a list of causes or triggers for the financial crisis. These topics serve as queries for our retrieval system to obtain potentially relevant doc-

[1]The section titles, such as *subprime lending* and *growth housing bubble*, from “Financial Crisis of 2007-08” Wikipedia page form the set of topics we consider.

[2]Available at http://ir.dcs.gla.ac.uk/terrier/.

[3]See http://en.wikipedia.org/wiki/Financial_crisis_of_2007.

uments. For each section title of the wikipedia page, such as "subprime lending", we extracted the text associated with each heading and used the Yahoo! Content Analysis API[4] to acquire important key phrases. We used both the title and the extracted key phrases from the textual content to formulate a total of 13 queries (See Table 2 for examples[5].). The extracted key phrases were added to the respective queries, because we anticipated that some of the documents discussing a given topic may not contain the title of the topic. Having the key phrases in the query would help recognize relevant documents in such cases. Lastly, we used the BM25 retrieval model (Robertson and Zaragoza (2009)) to rank documents relevant to each query.

Original Query	Expanded Query
subprime lending	subprime lending loans finance "mortgage lenders" "intense competition" "market share" "mortgage originators" "market power"
growth housing bubble	"real estate" median household income american house price

Table 2: Query Formulation Examples

3 Topic Discovery

The objective of this portion of the analysis is to understand what topics are prominent for each hearing/ meeting or speaker. (Here, we ignore the Wikipedia-based queries.) It maybe useful, for example, for analysts to determine what were (a) the main topics in each hearing/meeting[6], (b) the topics emphasized by each speaker, and (c) the important hearings for each topic.

We use the Dirichlet Multinomial Regression LDA[7] (Mimno and McCallum (2012)), where each hearing/meeting or speaker is assigned an indicator variable. These indicator variables are metadata for each meeting/hearing. The model learns weights on these metadata and allows us to sort topics by importance for each hearing and speaker. We use these weights in the query focused summarization task described in section 4.

The raw text was processed to exclude stop words and infrequent words (words occuring less than 5 times).

We present results here from two sets of meetings - the FOMC meetings and the Dodd-Frank hearings (see Table 3).

	FOMC	Dodd-Frank
Speakers	56	271
Meetings	14	61
Speech Events	3,032	11,954

Table 3: Summary of Meetings

We illustrate our analysis using results from the Dodd-Frank Hearings. Our analysis modeled 200 topics for these hearings and one of the topics is related to executive pay [8]. Hearings "56241", "56767", "48873", "48875" and "54589" discussed the topic the most. Interestingly the first two hearings are both defined as hearings on executive pay in the financial sector.

Hearing "48873" is "a special hearing,an ad hoc hearing, called as the second half of our conversation about the question of bonuses paid to the AIG (...)". To investigate this further, we looked at a AIG centric topic [9] and hearing "48873" is the third most important hearing for this topic. This suggests that the model is at least intuitively consistent.

Table 4 show examples of the most important topic by person for the Dodd-Frank hearings i.e. the topics of "expertise" for each person.The assumption is that each individual is an expert in one topic. Importance is measured by the proportion of the speech of the individual taken up by a topic. The frequency of speech by the individual does not affect the importance measure. Table 5 presents important topics for Republicans and Democrats in these hearings.

Speaker	Topic
Mark Froeba Principal, PF2 Securities Evaluations	rating agencies credit ratings agency investors issuer investor issuers moody
William Francis Galvin, Secretary, The Commonwealth of Massachusetts	regulators agencies oversight regulating regulatory things coordination exist multiple entities

Table 4: Dodd-Frank Hearings - Topic by Person

[4] https://developer.yahoo.com/yql/console/

[5] For more example, please see http://www.cs.cornell.edu/~luwang/polinformatics/queryPhrases.txt.

[6] Sometimes these meetings can be fairly open-ended. E.g."Members are obviously free to raise anything they want today, but it is my hope that we would focus on these very important questions of financial regulation., Dodd-Frank/CHRG-110hhrg44900

[7] A manual approach is not easy to scale-up. Alternatively, simple approaches such as a simple word count will yield common phrases/ words and will not provide us with a summary of the topics of interest for each hearing/ speaker. Similarly relying on query match alone may yield high precision matches but reduce recall.

[8] The top 10 words are - "compensation executive incentives executives pay stock performance company incentive top"

[9] The top 10 words are - "aig company correct counterparties fp september taxpayer names billion foreign"

Similar output on all hearings /meetings and persons along with the topics themselves (tagged with the most important meetings for each topic) is available from the authors.

Republican	**Democrat**
fannie freddie mae mac government housing gses conservatorship put	time expired gentleman presiding recognize minutes chair gentlelady excuse illinois
guess question lot thought folks sort fine panel place comments	world country failure government responsible real continue called america capitalism

Table 5: Top 2 Topics for Republicans and Democrats

4 Query-Focused Summarization

In this section, we utilize summarization techniques to extract the salient information in the context of queries specified by the users. For example, a user might be interested in the policy change and the relevant discussions on "subprime lending" in different years.

In general, our summarization algorithm takes as input a set of documents (or top ranked snippets) and a user-specified query, then selects one sentence each time until a pre-defined length limit. For each iteration, we select the sentence based on scores output by three metrics – *SumBasic score* ($S_{SumBasic}$), *expertise score* (S_{expert}), and *Topic-relevance score* (S_{topic}). Specifically, our system uses the top 1000 snippets returned by our ranker in Section 2. We then describe how we compute each score below.

SumBasic Score. The first scorer we use to estimate the sentence importance is adopted from the SumBasic summarization system 22005Nenkova and Vanderwende), which measures the salience of each sentence based on word frequencies. Concretely, for each content word w_i, $P(w_i)$ is computed as $\frac{n_i}{N}$, where n_i is the frequency of word w_i and N is the total number of content words. The SumBasic score for sentence S is computed as:

$S_{SumBasic} = \frac{1}{|S|}\sum_{i=0}^{|S|-1} P(w_i)$

To encourage diversity, the probability of the word in the sentence selected in the previous iteration is updated as $P_{new}(w_i) = P_{old}(w_i) \cdot P_{old}(w_i)$.

Expertise Score. The expertise score is designed to encode the speaker role information in the summary. Intuitively, we would like to detect the people who play important roles in the policy making process or leading discussions. Given that the speaker identity is available for congressional hearings or FOMC meetings, we are able to model the contribution of each participant in different topics. Therefore, we obtain the word-topic distributions and author-topic weights from Section 3, and utilize those to estimate the expertise score for each sentence.

Step 1. For each word w_i, select topic t_i where $t_i = \max_t P(w_i|t)$.

Step 2. For each speaker s_j, obtain the weights on the topics from Section 3, i.e. $\theta_{t_k;s_j}$, where t_k indicates the kth topic.

Step 3. For each sentence S uttered by speaker s, compute the expertise score as $S_{expert} = \frac{1}{|S|}\sum_{i=0}^{|S|-1} \theta_{t_i;s}$, where t_i is the topic selected for w_i from step 1.

For sentences without speaker information, e.g. the ones extracted from congressional report, we use the default weights for θ_{t_j}.

Topic-Relevance Score. The topic-relevance score aims to capture the relevance of the latent topic structure of a sentence to the latent topic structure of a set of documents. Given a set of query-relevant documents, we apply a non-parametric mixture model – Hierarchical Dirichlet Process (HDP) (Teh et al. (2004)) to discover the latent topics in these documents. Specifically, we model each document as a group of sentences, and each sentence as "bag of words". The words in a sentence is generated from a number of latent topics, where a topic is modeled as a multinomial distribution on words. The topics are shared among sentences from different documents.

We identify sentences that best represent a document by comparing the similarity between the topic proportions of a sentence and the document. Here are some sample word clusters for two selected topics. "*house Household homes programs business large Or regulators meeting gross margin directly dangerous separate types Lending reversed fees products loaned*" is likely to talk about housing, and "*securitizing Securities determination Reflecting accuracy conglomerate Understanding Board unclear joint Resolution Federal subsequently provision competition 1980s Commodities Hearing Lehman influence*" may be about regulation.

Full Scorer. Given the three metrics, the final

CHRG/Dodd-Frank/CHRG-111hhrg53238	Mr. Chairman	MBA looks forward to working with the committee on new consumer protection and regulatory modernization legislation as these proposals develop.
CHRG/Other/CHRG-111shrg57319	Mr. Schneider	I think that is primarily true because Long Beach tended to originate higher credit risk assets than other subprime mortgage originators.
CHRG/Other/CHRG-111shrg57319	Mr. Beck	you send an email early in the morning , 7:17 a.m. Subject , Re Option ARM Delinquency to Ms. Feltgen and to Mr. Schneider, making a plan to supply loan-level detail and coordinate with finance.
CHRG/TARP/CHRG-110shrg50416		The interest rate charged will not be greater than the current Freddie Mac Weekly Survey Rate at the time of modification.
CHRG/Other/CHRG-111shrg57319		Another key component of WaMu's higher-risk strategy involved efforts to increase the company 's exposure to the subprime market.

Table 6: Sample summary for expanded query "subprime lending" by using all the documents.

2005		
FOMC20050630	MS. YELLEN	So Fannie 's and Freddie 's books may look better in some sense – less risky – than they really are because of all of the second mortgages going up to possibly 125 percent . "
FOMC20050630	MS. YELLEN	We 've had changes in the rules for tax exemption and in 1997 on capital gains from the sale of primary residences that would make holding real estate assets more attractive .
FOMC20050630	MS. YELLEN	One view that I think is very prevalent is that the use of credit in the form of piggyback loans , interest-only mortgages , option ARMs (adjustable-rate mortgages) , and so forth , involves financial innovations that are feeding a kind of unsustainable bubble .
FOMC20050630	MS. YELLEN	One of the things that we looked at that we thought was interesting was the behavior of price-rent ratios for residential housing and for commercial office space .
2007		
FOMC20071211	MS. YELLEN	CDS spreads from major financial institutions with significant mortgage exposure , including Freddie and Fannie , have risen appreciably .
FOMC20071211	MS. YELLEN	Banks are showing increasing concern that their capital ratios will become binding and are tightening credit terms and conditions .
FOMC20070321	MS. MINEHAN	Indeed , I spoke to members of the advisory board of Harvard 's Joint Center for Housing Studies in late February .
FOMC20070321	MR. FISHER	Bill , you talked about subprime mortgages in some detail but not about alt-A mortgages in great detail .
2008		
CHRG/TARP/CHRG-110shrg50415		The Chairman of the Federal Reserve , Ben Bernanke , and Treasury Secretary Hank Paulson and many other respected individuals have all agreed on that fact .
CHRG/TARP/CHRG-110shrg50417		If there were not Federal deposit insurance and access to the Federal-backed liquidity windows at the Federal Reserve and Federal Home Loan Bank , not a single one of these banks would have survived from August 2007 until today .
CHRG/Dodd-Frank/CHRG-110hhrg45625		At the Federal Reserve , we have sought to address financial market stresses with as minimal exposure for the U.S. taxpayer as possible .
CHRG/TARP/CHRG-110shrg50416		I will begin by talking about our activities as the regulator of Fannie Mae , Freddie Mac , and the Federal Home Loan Banks , and then turn to TARP .

Table 7: Sample summary for expanded query "subprime lending" for each year.

score for each sentence is given by:

$$S = \alpha \cdot S_{SumBasic} + \beta \cdot S_{expert} + \gamma \cdot S_{topic}$$

where $\alpha+\beta+\gamma = 1$. Given that the metrics may have different scale, we transform the the scores onto $[0, 1]$ in each iteration.

Sample Summary. Here we display the sample system summaries for the expanded query "subprime lending". Two types of summaries are demonstrated below: (1) a summary generated from all the documents (Table 6); (2) one summary per year (Table 7).

There are several observations based on the generated summaries. In terms of the resource, we find that congressional hearings usually provide the most important information when the summary is generated based on all the documents. When we generate summaries for each year, FOMC transcripts have significant contribution for most of the years. Moreover, by adding expertise score, we are able to obtain more information from important or influential speakers based on speaker roles. For example, Mr. Yellen play an important role on subprime lending only in year 2005 (Table 7), but not other years.

5 Conclusion and Future Directions

We develop a system to retrieve top relevant information based on user-specific queries, analyze topic structures with or without speaker roles, and generate query-focused summaries by considering content importance, expertise of speakers and topic relevance. Based on the summaries generated for different years, we can track the content change, and detect important documents or meeting transcripts, and speakers who play important roles for various topics.

In future work, we would like to incorporate knowledge from domain experts to improve our summarization system. For example, including more domain-specific preference scores in the selection of summary sentences. We would also like to collect feedback from domain experts on the output of our system and fine tune the system to produce useful summaries.

References

David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*, 2005.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, 2004.