**Objective**

Dialog Corpus Collection

Summaries with Pyramid Annotation

Identify central propositions /abstract objects under discussion within a topic .For Abortion they could be on similar lines

* **Women protest Anti-Abortion Bills**
* **Crime-abortion**
* **Poverty -abortion**

Use these propositions for computerized summary of dialogs with pros and cons of each proposition

**Data collection**

Scrape 4Forums website, it has support for quoting another post, helpful for creating dialogs with turns.

To begin with:

Collected dialog exchanges from topics Abortion and Evolution (selected these 2 as number of posts and subtopics was large).These dialog exchanges have more than 7 turns per dialog.

Samples file for a subtopic “most-id supporters not creationist” within topic “creation-intelligent-design evolution” is (**most-id supporters-not-creationistpairall**.csv)

To begin with the initial data collection, selected first 10 pages from each topic (Evolution and Abortion) each page had a list of subtopics thus total subtopics selected within each topic were 34.

**A close look into Data**

Some Authors post very frequently, it would be better to have posts across various authors giving a wider view.

Some of the posts were too long, going up to 500 words or even sometimes more. For summarization with 8 turns, need an upper limit on word count.

A few Dialogs continue up to 20 turns with same speakers in a sequence but majority of the dialogs end near to 12 turns. Could be a characteristic of an Author.

To get some intuition on these, grouped dialogs based on

* No of Turns in each dialog
* Maximum length of post in a dialog
* Author

(Sample files in :

**Abortion\_debate\_Analysis\_Author.xls, Abortion\_debate\_Analysis\_Turns.xls, Abortion\_debate\_Analysis\_Length.xls**)

Varying these may affect quality of dialogs selected for the task.

**A pilot with topic modelling**

Criteria for dialog selection:

At least 3 dialog turns within a subtopic

This resulted in 2 subtopics for evolution and 4 for Abortion.

* most-id-supporters-not-creationists
* evolution-religion
* whether-you-agree-not-its-dark-truth
* every-woman-has-right-survive-pregnancy
* women-protest-anti-abortion-bills-virginia
* why-pro-life.html

Haven’t Analyzed results much except that a few key words for subtopic “most-id-supporters-not-creationists “could be helpful.

Result files in drive

**Evolution\_composition.txt**

**Abortion\_composition.txt**

This is all what I have till now… along with agreement disagreement.

I need to start working on the document itself now as I haven’t had much progress on that front

An increasing amount of information and dialog exchange occurs in natural dialogue in online forums, where people share their opinions about a vast range of topics. In particular, there is a lot of disagreement in these forums and I am interested in identifying the facets and propositions central to these debates and disagreements and summarizing these facets along with their pros and cons.

The research aims to build a Dialog Corpus including sample Dialog Summaries with Pyramid annotations using crowdsourcing .These could help identify the properties for future dialog summarization task. Explicit Dialog behaviors, discourse cues marking relations of comparison and contrast may be useful to determine the propositions central to a dialog.

The end goal would be to build a short summary containing the central propositions or facets in the dialog and identify pros and cons of each .