ACM - ATTRIBUTE CONDITIONING FOR ABSTRACTIVE MULTI DOCUMENT SUMMARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Abstractive multi document summarization has evolved as a task through the basic sequence to sequence approaches to transformer and graph based techniques. Each of these approaches has primarily focused on the issues of multi document information synthesis and attention based approaches to extract salient information. A challenge that arises with multi document summarization which is not prevalent in single document summarization is the need to effectively summarize multiple documents that might have conflicting polarity, sentiment or information about a given topic. In this paper we propose ACM, attribute conditioned multi document summarization, a model that incorporates an attribute conditioning module in order to decouple conflicting information by conditioning for a certain attribute in the output summary. This approach shows strong gains in ROGUE score over baseline multi document summarization approaches and shows gains in fluency and informativeness as shown through a human annotation analysis study.

1 INTRODUCTION

Abstractive multi document summarization is the task of writing a single summary of the key points and content in multiple related documents. This task has evolved from research in single document abstractive and extractive summarization however it faces unique challenges due to the duplicate content, conflicting content, a larger body of text as well as inter document connections between ideas. ((**?**) This task has evolved from early approaches using sequence to sequence (Seq2Seq) neural architectures to transformer based architectures and the introduction of large-scale datasets ((**?**) (**?**)). Beyond the introduction of approaches now commonly used for single document abstractive summarization, cross document attention and graphs that capture relations between text in various documents have further improved the state of the art for multi document summarization tasks. ((**?**), (**?**)). These graphs aim to better represent the inter dependencies between articles by representing text spans as nodes in the graph and capturing the relations between these sentences as edge weights.

Despite the advances made with these approaches, a significant challenge remains with multi document summarization with respect to how to deal with contradictory information present in the multiple source documents. It is critical to both learn the relationships between different documents as well as to extract salient information that is consistent with the output viewpoint. This is a situation often faced with summarizing multiple news articles where different viewpoints on an issue can significantly change the semantic structure of the content present in each article making it challenging for the abstractive summarization model to learn the relationships between inconsistent or conflicting information.

This paper proposes ACM, attribute conditioned multi document summarization, a novel approach that incorporates an attribute conditioning module with abstractive multi document summarization in order to condition for a particular attribute when generating the multi document summary. This approach addresses the challenge of dealing with conflicting information in the input documents by conditioning for a particular attribute in the input text. We further analyze the contributions of conditioning with this attribute model by using a weighting term to condition for the attribute when learning a graphical representation of the input documents, during train time and during evaluation. The attribute conditioning model is trained in order to determine view point consistency through sentiment and polarity classification datasets, however this module is highly composable and can be conditioned for other attributes as well. We train these classification models on every input prefix

Source 1
President Donald Trump quietly signed a bill into law Tuesday rolling back an Obama-era regulation that made it harder for people with mental illnesses to purchase a gun. The rule, which was..
Source 2
President Donald Trump signed a measure nixing a regulation aimed at keeping guns out of the hands of some severely mentally ill people. The original rule was...

Table 1: Contradictory Article Content

in the input dataset in order to train the model to be agnostic to input text length. We evaluate these approaches on baseline abstractive multi document summarization architectures in order to observe improvements in the output consistency evaluated through ROGUE metrics and through human annotations for fluency, informativeness and consistency. Our approach consists of individual composable elements, each of which we further evaluate independently through ablation studies.

- We learn a graphical representation of the input documents that weights graph edges incorporating both the conditional attribute score for each input as well as the cosine similarity.
- We conditionally fine tune the abstractive multi document summarization module by combining the logits for each input prefix passed to the decoder module with the conditioning score for that prefix from the attribute conditioning module.
- When evaluating the model, we modify beam search to rank each beam according to the product of the attribute conditioning model and the conditional likelihood score.

The contributions of our work are as follows:

- We train an attribute conditioning MDS model to address conflicting information in input documents and show that our architecture combining this module with abstractive multi document summarization improves ROGUE metrics for the output summary.
- We provide a human annotation analysis on the output summaries evaluating for informativeness, consistency and fluency.
- We analyze the contributions of each composable element in our model.

2 Related Work

This paper develops on techniques developed through the following NLP tasks - abstractive summarization, multi document summarization, and conditional language modeling - to efficiently address the issue of decoupling conflicting attributes in multi document summarization. We define the task of incorporating conflicting attributes e.g. sentiment and polarity as defined by a classification model in abstractive summaries as an application of conditional language modeling techniques to condition for a particular attribute when generating the output summary thus removing conflicting information. Conditional language modeling is a key factor in this problem as we aim to decouple the conflicting information in the summaries by conditionally selecting which text output to work with.

2.1 Conditional Language Modeling

Conditional language modeling has been approached both through applying global constraints on text generation, by applying control only at inference time and by directly optimizing through policy gradient methods. (**?**), (**?**), (**?**). Alternative approaches include relying on predefined sets of control tokens or control codes as a form of a copy mechanism (**?**). These approaches have been successful at steering language models towards specific features or as in (**?**) with conditioning for the outputs to include words conditioned for through a simple bag of words attribute model. We build on these approaches to design the attribute conditioning module.

2.2 Abstractive Summarization

Abstractive summarization (AS) is the process by which a language model is trained to best write original text that matches a pre-generated summary for a given article. AS has gone through several phases of which the pioneering work was carried out by (**?**) where an Seq2Seq RNN Model was implemented to generate text. The Seq2Seq RNN Model inherently had multiple challenges such as altering factual details and redundancy. (**?**) circumvented the issues by creating a Pointer-Generator Model which keeps track of words and sequence in the original text and using them in the result hence ensuring the meaning of summary is in-line with the original text. This paper also included a coverage mechanism to keep of track of which parts of the original text have been summarized thus penalizing repetition. This work was built on through the development of BART, the bi-directional auto-regressive transformer. (**?**). BART improves on the task of abstractive summarization by introducing arbitrary noise in the input text and training the model to reconstruct the original text.

2.3 Multi Document Summarization

Multi document summarization has evolved through four phases of primary approaches since the task was first introduced. The first set of approaches focused on graph ranking based extractive methods through TextRank (**?**), LexRank (**?**)and others. These approaches came before syntax and structure based compressive methods which aimed to tackle issues of information redundancy and paraphrasing between multiple documents. Compression-based methods as shown in (**?**) and paraphrasing based (**?**) (Bing et al 2015) were improved upon with the advent of neural seq2seq based abstractive methods in 2017. This allowed multi document summarization to further improve upon the work done with single document abstractive summarization through approaches such as pointer generator- maximal marignal relevance (**?**), T-DMCA (**?**) the paper that also introduced the foundational WikiSum dataset and HierMMR (**?**) that introduced MultiNews. These approaches aimed to tackle information compression through maximal marginal relevance scores across documents and through attention based mechanisms.

Improvements since those baseline models include further leveraging graph based approaches to pre-synthesize dependencies between the articles prior to the multi document summarization as tackled in (**?**). Further work needs to be done to further exploit these graphical representations as (**?**) essentially works to establish baselines with tf-idf, cosine similarity and a graphical representation first described in (**?**). These papers primarily aim to address de-duplicating information and learning relationships between the different topics shared across documents. The first paper to work with addressing conflicting information across the multiple documents is (**?**) in their work to incorporate an opinion polarity module to the pointer generator network architecture used earlier on in multi document summarization research.

3 Our Approach

We present a novel technique ACM, attribute conditioned multi document summarization, which is designed to address the problem of resolving conflicting information in multi document summarization through the use of an attribute conditioning module. Conflicting information is determined by the attribute conditioning model trained on both a polarization and a sentiment analysis dataset since both factor into determining information consistency. XLNet (**?**) is used as the model architecture for the attribute conditioning module and is trained using sentence prefixes in order to capture both word level and phrase or sentence level features. XLNet has shown state of the art results in sentiment analysis tasks. The outputs of this classifier are used in each approach in order to fine tune the model to consistently condition for a particular attribute.

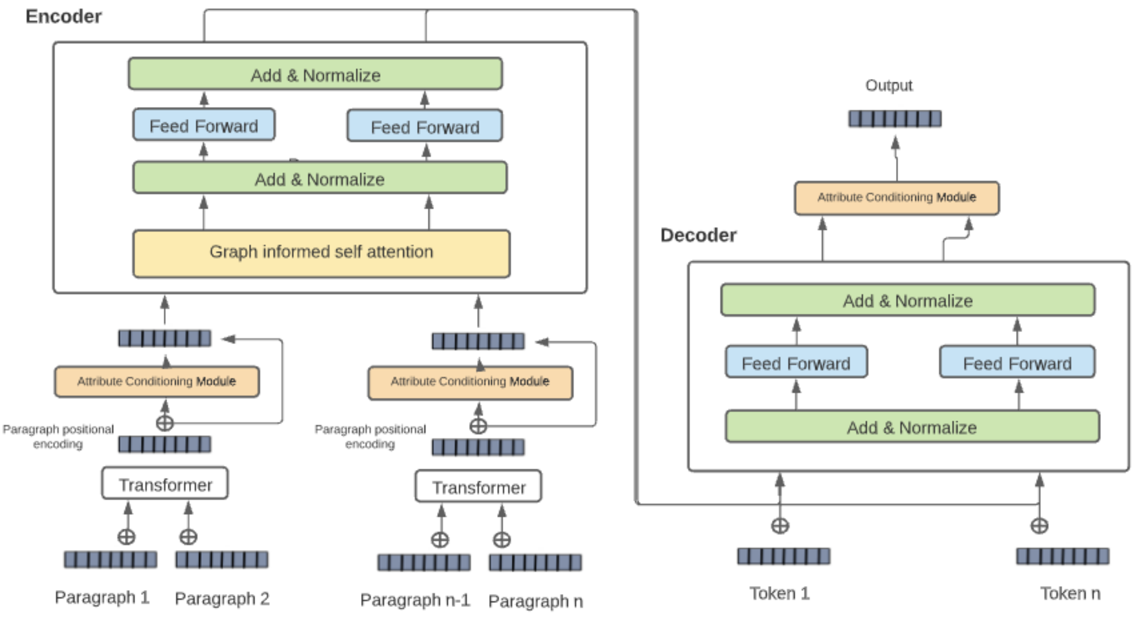

Figure 1: Attribute conditioned multi document summarization (ACM) model diagram. At each stage of the summarization process, ACM uses the attribute conditioning module to preserve viewpoint consistency in the output summary.

In order to preserve viewpoint consistency, the attribute conditioning module is used to guide each stage of the summarization process. To prevent over fitting to the conditioning model, each stage is conditioned with a weighting term set as a hyper parameter when training ACM. Each of these approaches is explained in more detail in the following sections. We also include each approach as an ablation study in order to determine how effective each stage of conditioning is on the final output summaries. In these ablation studies, we aim to investigate how well the attribute conditioning model can show improvements over three different baseline abstractive multi document summarization models - BART (**?**), BART with longformer self attention (**?**) and graph sum (**?**).

3.1 Graph conditional weighting layer

We preprocess the inputs to the model through a graph conditional weighting layer. As shown in diagram 2, for each of the input paragraph vectors both the positional encoding layer and the attribute conditioning layer outputs are computed. The graph representation is then constructed by creating a matrix of values where each row and column represents a sentence in one of the input documents. The baseline model for GraphSum (**?**) computed a similarity score between these paragraphs by using tf-idf. In our proposal, we aim to improve this approach by also incorporating the similarity between the attribute scores between each paragraph. Thus the graph will learn stronger weights between paragraphs that are of the same polarity or sentiment and conditionally select for those when generating the output summary.

After generation, this graph is passed into the transformer encoder matrix as a set of attention weights for the graph encoder layers in combination with the standard transformer encoder layers in the model. Each sentence is passed through the conditional model and the score is multiplied and normalized before generating the matrix of weights for the transformer encoder layers.

3.2 Attribute future discriminators

We treat the problem through a Bayesian factorization approach effectively decoupling the pretrained summarization model from the fine tuning classification model. This approach is described as using a future discriminator as it takes the entire sequence of text generated so far, appends to it the most likely next token predicted by the summarization model and then computes the likelihood that this sequence satisfies the desired classification attribute. In doing so it increases the likelihood of both generating text that satisfies the summarization model with a high likelihood as well as staying consistent with the desired text modeling attributes.

We can model the standard summarization task as

$$P(X) = \Pi_i^n P(x_i|x_{1:i-1})$$

Let X refer to a set of input documents $d_1, ..d_n$ pertaining to a specific MDS summary s_i. Let each document $d_i = x_{i1}, ..x_{in}$ where x_{ij} refers to token j in document i. Prior to the MDS task, all input documents are appended, thus absolving the need to index each token with its corresponding document. Lastly, let a refer to the conditioning attribute. When conditioning on a certain attribute, we model the task as

$$P(X|a) = \Pi_i^n P(x_{1:i-1}, a)$$

We rely on the following factorization to decouple the summarization module from the attribute conditioning model:

$$P(X|a) \propto P(a|x_{1:i})P(x_i|x_{1:i-1})$$

This approach allows us to train the conditional model and the summarization model separately and in a composable manner to achieve the desired output conditioning.

```
Result: Conditional MDS with Future Discriminators
beams = []
 beamWidth = 200
 scores = [];
 for t = 1...T do
    topBeams = argmax(beams, beamWidth);
     for $beam \in topBeams$ do
        topNextTokenLogits = argmax(nextTokenLogits, 200)
         topNextTokens = vocabToIndex(topNextTokenLogits)
         combinedScores = []

        for token, tokenLogits $\in (topNextTokens, topNextTokenLogits)$ do
            beam' = beam + token
             conditionalLogits = conditionalModel(beam')
             combinedScores += conditionalLogits*topNextTokenLogits
             beams = beams $\bigcup beam'$
             end
        end
    end
```

Figure 1: Future Discriminators for MDS Beam Search incorporates the attribute conditioning model at evaluation time in order to weight each beam based on the combined attribute score and maximum log likelihood.

As both the summarization model and the classification models are pretrained, during evaluation, each of the top k words selected by the decoder logits in the summarization model is added to the best sequence so far and then passed into the attribute classification model. The output probabilities of both models are then combined in selecting the next best word generated by the combined model. Technically this allows for a high degree of composability as each of the attribute models can be layered on top of each other and be added to the final logits with different weights. This shows the Bayesian factorization approach used for conditionally generating text based on the selected attribute classification model.

3.3 Conditional training

At each time step of training, we combine the probability distribution over the next words generated by the decoder with the polarity scores as determined by the attribute classification model. In contrast to other models such as (**?**), this allows the base model to be conditioned to output based on

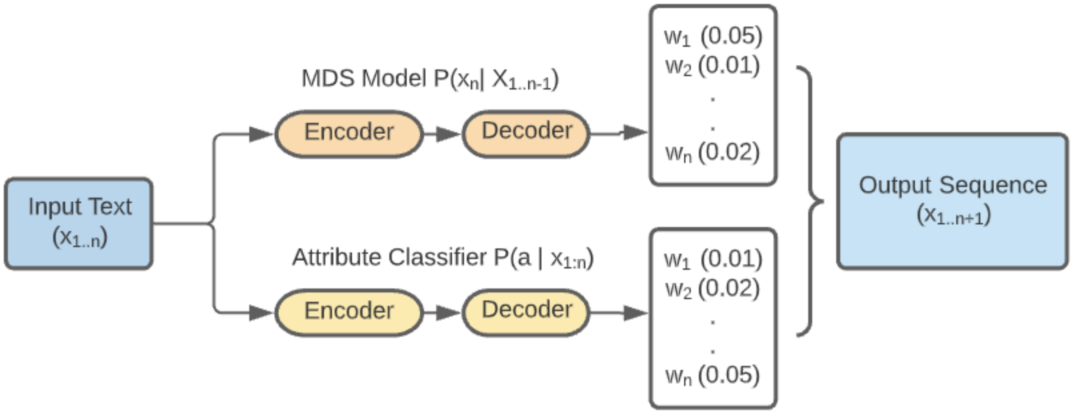

Figure 2: Conditional MDS diagram

the desired attribute similar to the approaches taken to train class conditional language models. An additional step is taken to A diagram of this model can be seen in Fig. 2.

For each model architecture, the decoder logits are multiplied with the attribute model's logits during during training. This trains the abstractive multi document summarization model to output text conditioned on that particular attribute. The same classifier models and baseline MDS models were used to evaluate this approach as in the previous method. We show that this approach can be applied to any baseline MDS architecture in the follow up ablations.

4 Experiments

4.1 Experimental Setup

Classification Models We first train an XLNet (**?**) attribute classification model on two different datasets AllTheNews (**?**) and the MPQA Opinion Corpus dataset (**?**) in order to determine the sentiment and polarity scores of input text. AllTheNews consists of a body of In order to create a dataset most similar to the input sequences that would be passed through the future discriminator approach as well as the conditionally MDS models, the input text in the dataset was augmented to include all prefix length subsequences with the same label. This corresponds to the prefix length subsequences that will be evaluated for the above approaches. AllTheNews dataset consists of 2.7 million articles from 26 different publications ranging from January 2016 to April 2020 in English. (**?**). This dataset was augmented with polarity labels according to the news source label. The MPQA Opinion Corpus was chosen over other sentiment analysis datasets as it consists of articles pulled from news articles on a broad range of news sources and is consistent with the approach used in (**?**)

MultiNews (**?**) consists of a varied set of news articles spanning over 1500 sites and their corresponding human written summaries. The average length of the source documents is 2100 tokens which allows the Bart+Longformer model trained with input length 4096 tokens to consume multiple concatenated articles at a time. For all models, each of the input documents are concatenated together and fed into into the model in batches of maximum token size.

Training Configurations ACM was trained using with 8 transformer encoder heads and 6 graph decoding layers. Beam size was set to 5 with length penalty factor 0.6 trained with gradient accumulation every 4 steps. Additionally we ran a hyper parameter search to determine the ideal weighting terms for the attribute conditioning module at each phase of summarization. For the ablation studies, BART and BART+Longformer self attention models were trained using the same set of hyperparameters used to train the baseline models. All models were trained using 2 NVIDIA K-90 GPUs. Beam search for each approach set a length penalty of 2.0, max length of 200 tokens, with 4 beams. A BART architecture with 8 transformer heads and 6 decoder layers was used with gradient accumulation every 4 steps. The BART model was trained with max token size of 512 and the Bart+Longformer self attention model was trained with max token size of 4096. GraphSum (**?**) was trained using the same configurations present in the original paper in order to reproduce results with 8 transformer encoder heads and 6 graph decoding layers. To maintain consistency with ACM, beam size was set to 5 with length penalty factor 0.6 trained with gradient accumulation every 4 steps.

5 Results

We evaluate our model on both ROUGE scores as well as perform human annotations to evaluate output summaries for fluency, informativeness, and consistency. (**?**). In addition we perform a series of ablation studies for each technique presented in order to assess the contribution of each to the final result.

5.1 Evaluation Results

We primarily compare ACM against other strong baseline model architectures including BART (**?**), BART+Longformer attention (**?**) and GraphSum (**?**).

Model	**Rouge-1**	**Rouge-2**	**Rouge-L**
BART	48.54*	18.56*	20.84*
BART+Longformer	49.03*	19.04*	24.0*
GraphSum	42.99*	27.83*	36.97*
ACM with Sentiment	50.05	27.98	38.10
ACM with Polarity	**50.12**	**28.12**	**38.19**

Table 2: Evaluation results for MDS graph with conditional weighting evaluated on the MultiNews dataset with ROUGE scores. Stared numbers are reproduced from the original papers.

We evaluate each of these methods on the MultiNews dataset. The summarization quality is measured with standard quality metrics ROUGE-1 (overlap of unigrams), ROUGE-2 (overlap of bigrams) and ROUGE-L (longest common subsequence) between the generated summaries and the gold standard references. ROUGE-L is often used as a measure of accessing fluency. (**?**). We include an appendix with comparisons between the summaries generated by each model and their corresponding gold standard summaries. The first block of results represents the baseline models' ROUGE scores followed by the results from ACM. Overall performance shows that the sentiment attribute module and the polarity attribute module perform about on par with polarity performing slightly better. This trend holds across the other approaches as well. GraphSum with sentiment outperformed the BART based models as well confirming our hypothesis that learning the relationships between the input data improves attribute consistency in the final output summary.

5.2 Human Evaluation

In addition to the automatic evaluation reported above, we also perform a large scale human evaluation study on Amazon Mechanical Turk on the summary outputs. We randomly select 182 input test summaries from the MultiNews dataset and the corresponding output summaries generated by ACM. In order to assess the quality of the model irrespective of the classifier model chosen, we randomly selected output summaries between the two classifiers. We use the ACM model without conditioning as the baseline model. Annotators assess the overall quality of the summaries based on three different criteria: (1) informativeness, (2) fluency and (3) repetitive content. Informativeness is defined as the number of unique facts / pieces of information present in the summary. Fluency is defined as the readability of the text accounting for good grammar, noun phrases and logical flow of information. Repetitive content comes from repeated words, phrases, or ideas throughout the output summary. Each of these attributes were assessed on a scale of 1 (worst) to 5 (best). According to this scale, a high score is preferable for fluency and informativeness and a lower score is preferable for repetitive content.

Model	Fluency	Repetitive content	Informativeness
ACM w/ conditioning	**3.91**	**1.74**	**3.86**
ACM Baseline	3.48	2.3	3.24

Table 3: Evaluation results from a human annotation study over 182 randomly selected output MDS summaries shows strong improvements over the baseline model with respect to fluency, informativeness and repetitive content on a scale of 1 to 5.

5.3 Model Analysis and Ablations

In order to determine the contribution of each method used within the ACM model, we performed additional ablations and model analysis. The key ablation studies included evaluating each approach on one of the baseline models, BART and BART + Longformer. We analyzed the results here both in terms of achieving higher ROUGE scores as well as maintaining information consistency. Table 5 presents the ROUGE scores for each of the approaches. We note that there are improvements from each approach individually with respect to the ROUGE score with attribute future discriminators performing marginally better than the other approaches. BART+Longformer achieves overall better performance on ROUGE scores as compared to BART primarily due to the longer input sequence lengths passed in. In addition to evaluating on ROUGE, an analysis done on how well each approach was able to preserve the overall attribute conditioning.

Baseline Model	Rouge-1	Rouge-2	Rouge-L
BART	48.54	18.56	20.84
BART+Longformer	49.03	19.04	24.0
Our Approach			
BART w/ graph conditional weighting	48.61	19.06	21.01
BART w/ conditional training	49.10	19.52	20.94
BART w/ attribute future discriminators	49.14	19.63	20.94
BART+Longformer w/ graph conditional weighting	49.32	19.34	24.42
BART+Longformer w/ conditional training	49.72	19.55	**24.52**
BART+Longformer w/ attribute future discriminators	**49.83**	19.85	24.32

Table 4: Ablation study comparing each of the approaches against baseline BART and BART+Longformer models

Additionally we conducted a qualitative analysis of how well each approach was able to condition for polarity and sentiment in the output summaries by evaluating the summaries through the trained attribute conditioning module as shown in Table 5. This shows strong out of domain analysis results as the attribute conditioning module for polarity was trained on a different dataset, AllTheNews, and evaluated on the MultiNews dataset for MDS. This ablation study shows that This analysis showed that MDS with future discriminators is the strongest attribute conditioning model. Since the sentiment and polarity of articles as determined by the XlNet classifiers are used to compute the opinion of an article, we used these models to analyze the MultiNews dataset.

Approach	Mean	Std Dev
Graph conditional weighting layer	0.76	0.288
Conditional training	0.82	0.274
Attribute future discriminators	**0.891**	**0.103**

Table 5: MDS with attribute future discriminators shows the highest mean polarity score and narrowest standard deviation.

6 Conclusion

In this work, we present a novel approach ACM, attribute conditioned multi document summarization, that sets the new state of the art for multi document summarization. It tackles the challenge of addressing conflicting information in multi document summarization by conditioning for a desired attribute and preserving consistency in the final output summary. Through the attribute future discriminators we are able to compose different conditional attribute models with a pretrained MDS model during evaluation. To our knowledge, this is the first approach taken to effectively decouple conflicting information by conditioning for a certain attribute in the output summary. This approach shows strong gains in ROGUE score over baseline multi document summarization approaches and shows gains in fluency and informativeness as well as a reduction in repetitiveness as shown through a human annotation analysis study.

A First Appendix

Selected prediction samples for baseline model against each experiment.

A.1 Example 1

BART Baseline – The Siskiyou County Board of Supervisors voted 4-1 Tuesday to support the state's split from California, the Sacramento Times-Standard reports. The move would bring representation to rural North State counties that are beholden to the whims of representatives of the more heavily populated Southern California and free them from burdensome state regulations, the Times-Standard reports. "We have to have government that's local, understands our issues and has empathy," says a Scott Valley rancher who is leading the charge to form a new state.

ACM conditioned with Polarity
Not a big fan of Southern California? Neither is Northern California, apparently. Supervisors in rural Siskiyou County voted 4-1 Tuesday in favor of seceding from the state, reports the Times-Standard. The county thinks state officials in Sacramento are too focused on the big metro areas of the south and thus want to form a state of its own—to be called Jefferson. Supporters want to invite other northern counties, and perhaps a few from Oregon, to join, too, notes the Los Angeles Times. "We have to have government that's local, understands our issues, and has empathy," says one local rancher. It's clearly a very long-shot bid, one that would eventually require the blessing of both the state legislature and the US Congress. Still, neighboring Humboldt County is expected to meet soon to consider the idea. (Meanwhile, some counties in Colorado want to form a new state of their own.)

A.2 Example 2

BART Baseline – Microsoft's acquisition of Nokia is aimed at building a devices and services strategy, but the joint company won't take the same form as Apple, reports ZDNet. Microsoft has been working on its evolution into a devices and services company, away from the services business it traditionally was, for several years now with limited success. Its acquisition of most of Nokia is the latest acceleration of that strategy—to move further away from the moribund world of the beige desktop and towards the sunlit world of smartphones and tablets. Owning the desktop (via Windows) and building additional services on top.

ACM conditioned with polarity
Why did Microsoft buy Nokia's phone business? We now know Microsoft's answer: The computing giant released a 30-slide presentation today arguing that the move will improve Microsoft's margins on Windows phones, which will allow it to invest more in the platform, which will accelerate sales and market share growth, the Washington Post reports. But John Herrman at Buzzfeed has another explanation: "Fear of dying alone." Here's what he and other pundits are saying: The presentation "manages to sound both insane and uninspiring, outlining modest goals that still sound unrealistic," Herman argues—like capturing a whole 15% of the smartphone market. "It's a fitting end for the close of Microsoft's Ballmer era, during which the company … missed out on the most important change in consumer electronics in decades" while remaining profitable in unglamorous ways.

A.3 Example 3

BART + Longformer Baseline – The Supreme Court has a new term that could be a blow to the labor unions and roll back affirmative action at state universities, reports POLITICO. The justices are already facing a batch of petitions involving the rights of religious institutions to opt out of providing contraception under Obamacare. The case, which was brought by California schoolteacher Rebecca Friedrichs and other teachers, was brought by Orange County, Calif., schoolteacher Rebecca Friedrichs and other teachers, who are arguing that forcing government employees to pay union dues violates their First Amendment rights.

ACM conditioned with polarity The Supreme Court is facing a docket of high-profile political cases that will test whether recent liberal victories were more fluke or firm conviction, the New York Times reports. The court—which is divided 5-4 for conservatives, but saw Justice Roberts vote liberal on Obamacare and same-sex marriage—will look at cases including unions, affirmative action, and possibly abortion. A primer: Unions: Since 1977, unions have been allowed to charge non-union workers for dues that go to collective bargaining efforts, but not political ones. Now California teachers have brought a case saying collective bargaining is itself political. ”It could set the stage for a Citizens United-style reconsideration in the area of union dues,” a lawyer says. Affirmative Action: Abigail Fisher says that being white played a role in the University of Texas denying her admission back in 2008.

A.4 Example 4

BART + Longformer Baseline – A preservationist says he has found evidence that a Manhattan building is the former site of an 18th-century tavern where George Washington is believed to have enjoyed a celebratory drink during the American Revolution. If it is indeed the home of the legendary watering hole, the discovery could mean that the building that is perhaps Manhattan’s oldest is slated to demolished. ”After the English had marched up the Bowery and out of the city (in 1783), George Washington and Governor (George) Clinton stopped at the Bull’s Head (tavern).

ACM conditioned with sentiment - In 1783, after the British soldiers left New York City, George Washington is believed to have stopped for a celebratory drink at the Bull’s Head tavern. Now a preservationist thinks he’s found the historic site—and if he’s right, it could be the oldest building in Manhattan. Adam Woodward had heard that the building at 50 Bowery, currently scheduled to be demolished so a hotel can go up, might have ”the Bull’s Head’s structure, cellar, bones,” he tells CBS New York. So he searched the basement, and ”found myself in what I am pretty certain is the 1750s historic tavern,” he says. Specifically, he found what he thinks are hand-hewn and hand-planed joists and foundation walls from the Colonial era. Since that time, the building has housed a drugstore, a Chinese restaurant, and a beer garden, among other things.