Extractive Text Summarization of New Articles using Clustering and Sentence Relevance

Umang Shah
uks160030@utdallas.edu
Computer Science Department
The University of Texas at Dallas.
Richardson, TX.

***Abstract*—Generating summary of large pieces of text has been a common problem addressed by Natural Language Processing. The summary of any text document should present all or most of the important points in the text in the least number of sentences possible. News articles are a common piece of text which needs summarization. News articles provide detailed analysis of events, however, the summary of the article is enough to gain understanding of the news. In this report, a method of text summarization which uses 3 steps is explained. Sentence Similarity measures to identify how closely related each sentence is to the other. Clustering using these similarity measures to group these sentences by the information they convey. Finally, selecting sentences which are most informative within each cluster. Arranging the sentences by order of appearance in the original text a summary is formed.**

Keywords— summarization, clustering, news articles, natural language processing.

I. Introduction

News articles usually cover all the developments relating to the event being covered. However, due to constraints like time, we cannot read the entire article and would like to know just the important details, which are not conveyed by the headline. In such a case, a good summary of the . A good summary should be one which covers all different topics as well as views being expressed in the text. Many summarization techniques rely on selecting sentences which convey the most information. But it is possible that all these sentences convey the same information and thus leave out much of the information. The intuition of using clustering before sentence extraction is based on the fact that sentences which provide same or almost similar information need to be grouped to avoid all sentences of the same cluster being selected in the summary. So clusters of sentences are formed and each cluster covers the distinct points conveyed by the text.

The section 3 explains the approach and the motivation for selecting the approach. After that the implementation is detailed and section 4 explains the results that were obtained by running the implementation on raw newspaper articles. At the end the conclusion is provided which suggests improvements followed by references.

II. Related Work

Text Summarization has been a subject of great amount of research. However the specific method of clustering and sentence extraction has been demonstrated by Aliguliyev [1]. His work also advocates use of term similarities as a basis for establishing sentence similarity. The same intuition is used for formulating the sentence similarity measure for this technique.

Extensive research has also been conducted on sentence similarity measures. Achananuparp et. al. [2] evaluated 14 such measures. The tests conducted by them on 3 different data sets show that a combination of semantic and syntactic measure of sentence similarity have a better performance on average.

III. Approach

The approach used to generate summaries involves the following three key components:

A. Sentence Similarity Measures

A combined measure of semantic and unigram distance is used to derive a measure of sentence similarity.

1) GloVe Vector Cosine Similarity:

GloVe [3] is an unsupervised algorithm which gives vector representation of words developed at Stanford. Embedding words as a real valued vector they provide an effective method for measuring semantic similarity between terms. Building on the method of Aliguliyev, GloVe is used to derive semantic relation between terms and this is extended to sentence similarity.

For the purpose of implementing this project, pre trained glove vectors are used. The glove vectors have been generated by calculating co-occurrence over a large corpora of text (subset of Wikipedia 2014) and has vocabulary size of 400K words. The formula for calculating glove similarity uses cosine similarity of the computed vectors which appears in the semeval 2016 paper by Agirre et. al.[4]:

1. Summing vectors over all the words *w* in a sentence *s*.

$$v(s) = \sum_{w \in s} v(w)$$

2. Cosine Similarity between vectors of two sentences *s1, s2*.

$$gvSim(s1, s2) = \frac{v(s1)v(s2)}{\|v(s1)\|\|v(s2)\|}$$

2) Unigram Similarity:

As pre trained glove vectors are used, words like proper nouns, organization names etc. which appear commonly in news articles will not be present. To offset this difference Unigram probabilities are used to calculate surface lexical similarity as defined by Lin et. al. [4]

$$ugSim(s1, s2) = \frac{2 \times \sum_{w \in s1 \cap s2} \log P(w)}{\sum_{w \in s1} \log P(w) + \sum_{w \in s2} \log P(w)}$$

3) Combined Similarity:

Combining the glove and unigram similarity a combined measure is formed which takes weighted sum of the two quantities. The formula is similar to the one tested by Achananuparp et. al. for combined semantic and syntactic similarity. The combined similarity of glove and unigram probabilities is calculated as:

$$comboSim(s1, s2) = \alpha \times gvSim(s1, s2) + (1 - \alpha) \times ugSim(s1, s2)$$

* α is the weight ; taken as 0.7 in the project.

B. Sentence Clustering

Once distance measure is in place. A modified K-Means algorithm is used for clustering sentences based on their similarity to each other.

The Steps used by the modified K-Means are as follows:

1. Choose K initial sentences to represent each cluster. Initialization is done by selecting sentences most dissimilar from the headline.
2. Start with empty clusters and assign the representative sentence of each cluster.
3. Now consider all other sentences one by one and assign them to one of the clusters, whose representative sentence has maximum *comboSim*.
4. Re calculate the new representative sentence for each cluster by choosing the sentence which has highest average similarity with rest of the sentences.
5. Repeat steps 2-5 till convergence.

C. Representative Sentence Selection

Once clusters are efficiently created by the K-Means, the sentences are grouped by their information content as closely as possible. After that, the most important sentence from each cluster need to be selected to be part of the final summary.

1) TF-ISF:

Inverse Document Frequency is modified for the sentence level as Inverse Sentence Frequency using which TF-ISF is calculated. This measurement helps in measuring importance of terms present in the sentence.

Term Frequency, of each term *t* in sentence *s* in the article *A.*

$$tf(t, s) = \frac{f_{t,d}}{\sum_{t' \in A} f_{t',d}}$$

Inverse Sentence Frequency: nt is the number of sentences where the term t appears.

$$idf(t, A) = \log(\frac{N}{1 + n_t})$$

TF-ISF is calculated.

$$tf - isf(t, s, A) = tf(t, s) \times idf(t, A)$$

Sentence Selection Criteria: Summation of tf-isf of terms in the sentence normalized by number of words in the sentence.

$$ssc(s) = \frac{\sum_{t \in s} tf - isf(t, s, A)}{|t(s)|}$$

The sentence with highest value of ssc(s) in each cluster are selected to form the summary.

We take the summation of the tf-isf of terms and normalize it based on no of words in the sentence. The sentence in the cluster with the highest value is selected to represent the cluster.

2) Headline to Sentence Similarity.

Another approach is based on query based summarization approach. A news headline is meant to convey the idea of the entire articles in one line. This news headline serves as a good query to generate a summary for. By calculating the sentence closest to the query using the combined similarity measure. The sentences to represent each cluster are selected.

From each cluster select sentence with max comboDist w.r.t. the headline.

IV. IMPLEMENTATION

An implementation was developed in Python 3. nltk [5] is used to assist with word and sentence tokenization. System is broken down into modules and data flows from module to module to generate results. Two summaries are generated one using tf-isf and the other using distance sim with headline.

A) Software Architecture:

The software architecture is shown in figure. 1.

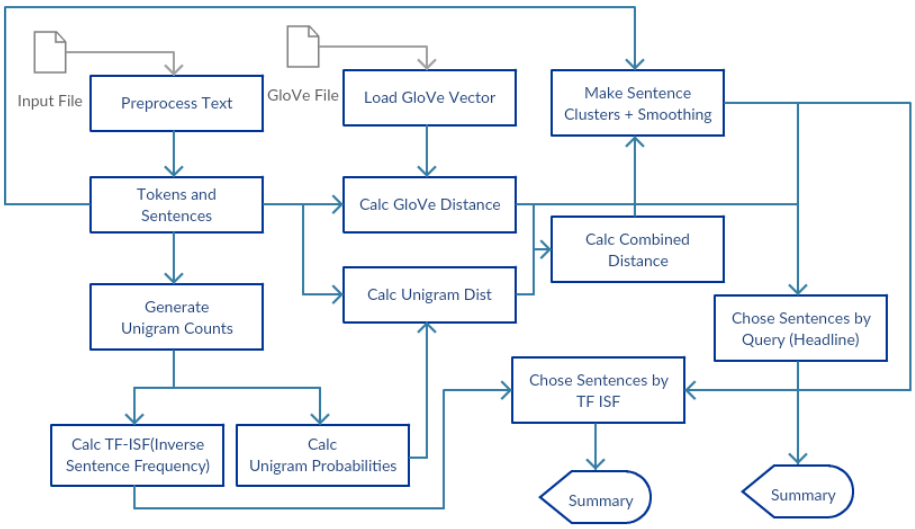

Figure 1. Software Architecture.

B) Main Flow:

1. The process begins with the input text file.
2. The first line has the headline and the article follows.
3. File is read and punctuations are stripped.
4. After cleaning and preprocessing the text, tokenization is done.
5. These tokenized word and sentence lists are used to generate the unigram counts of occurrence of each word.
6. After the initial steps, Unigram counts help generate Unigram probabilities and TS-ISF values for each word.
7. The pre trained glove file is loaded into the system and sentence glove similarity is calculated.
8. The glove and unigram probabilities are combined to form the final similarity metric.
9. The previously discussed modified K-Means is used for clustering the sentences.
10. Clusters are formed and the two sentence selection criteria are applied to generate two summaries both summaries with the title are the output.

V. RESULTS

A) Data:

Documents for using as input for the software were taken from the plain text news articles at csmonitor [7]. For the project, 5 text documents were manually downloaded and the implementation was run for these documents.

B) Result Analysis:

The figures 2, 3 and 4. Show the results of clustering and sentence selection on a document of running the implementation. Figure 2. Shows the result of sentence clustering. The clustering algorithm scans through all the sentences and finds relevant similarity among sentences all across the document and not just isolated to certain parts. This is also representative of how news articles have re occurring ideas about the main topic. Noticeably, some clusters are large in size compared to other clusters which could be a consequence of the cluster center having high similarity with a large no of sentences in the text. When sentences are too long, the similarity measure being used here may assign high similarity with shorter sentences as they may have their main idea being a part of the long sentence. Particularly, conversational sentences which have people making statements in news articles seem to be the cause of such irregularities. These sentences not only discuss the main topic bust also provide personal or background information as the person or organization's statements.

Figure 3. shows the result of sentence selection from the cluster using TS-ISF. Mainly, what is observed is sentences do convey important information, and to a certain extent convey distinct ideas however structure of the sentences is unconstrained and conservatory sentences are selected. The same problem with statement sentences is observed.

Figure 4. shows result of sentence selection from the cluster using Headline as a query for similarity. We see that the output in this approach has better output as compared to the previous approach. That is because we can avoid conservatory sentences in favor of declarative sentences which have more relevance to the headline.

VI. Conclusions

Some issues come to light after a complete implementation and experimental analysis. Co-reference resolution is an important problem as pronouns occurring in summary do not have the subject and it's an important task to replace pronouns with the correct corresponding entity. Another issue is that even though sentence similarity is measures for long sentences usually contain multiple ideas and clustering.

One of the main improvements could be in the clustering approach by using the lexical heads of terms to find the theme of each sentence. As for sentence selection measure, the same measures may perform well if better clusters are formed.

REFERENCES

[1] Aliguliyev, Ramiz M. "A new sentence similarity measure and sentence based extractive technique for automatic text summarization." Expert Systems with Applications 36.4 (2009): 7764-7772.

[2] Achananuparp, Palakorn, Xiaohua Hu, and Xiajiong Shen. "The evaluation of sentence similarity measures." International Conference on Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, 2008.

[3] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

[4] Agirre, Eneko, et al. "Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation." Proceedings of SemEval (2016): 497-511.

[5] Lin, Dekang. "An information-theoretic definition of similarity." ICML. Vol. 98. No. 1998. 1998.

[6] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[7] http://www.csmonitor.com/layout/set/text/textedition

Title: New report finds cleaner air for many, but not all.

More Americans have been breathing easy in recent years. On Wednesday, the American Lung Association (ALA) released its annual "State of the Air" report, which tracks air quality for the years 2013-2015, both nationally and for specific metrics, like cities and types of pollution. While some areas showed continued improvement and others deterioration, overall the report found that "the number of people exposed to unhealthy levels of air pollution dropped to more than 125 million people," down almost 25 percent from 166 million in the years covered in the previous report (2012 – 2014). The report credits this progress to the Clean Air Act, one of the US environmental movement's landmark achievements passed in 1970, and the scientists who study this issue agree. But in some parts of the country – especially California – environmental advocates see a need for smart policy decisions that will continue cleaning the air. "This [report] is just further evidence that our efforts to clean up the air are working, but that we have to push further to help improve the air ... for everybody in the country," Chris Cappa, a professor in the department of civil and environmental engineering at University of California–Davis, tells The Christian Science Monitor in a phone interview. Air pollution in such places as Beijing and New Delhi has made headlines in recent years, but the mid-20th century saw many American cities wrapped in a similar haze. In October 1948, the air over Donora, Penn., got so thick with smog from its steel and wire factory that the industrial town's fire department handed out oxygen tanks to ailing residents. The pollution killed 20 people and left half of the town's 14,000 residents recovering in hospitals or at home. A series of policies meant to prevent this disaster from happening again culminated in 1970 with the Clean Air Act's passage. Under the law, the US Environmental Protection Agency (EPA) must set science-based air quality standards for common types of air quality pollutants, and requires states to set enforceable limits on meeting those standards. Other components of the act cover interstate pollution and mobile emitters, such as car tailpipes. "While there are other factors" behind the drop, "the Clean Air Act has probably been the most important component," explains Cort Anastasio, professor in the air, land, and water resources department at UC-Davis, in an email to the Monitor. "EPA regulations," he continues, "while regularly derided by some in (and out) of politics, are making our air cleaner and provide benefits (in the form of increased human health) that far outweigh their costs." Contrary to concerns that these regulations would stifle businesses, the EPA estimates that just one provision of the act – 1990 amendments that target acid rain – will yield an economic benefit of $2 trillion by 2020, at an enforcement cost of $65 billion. Already, the ALA's report notes that America's gross domestic product has grown by 246 percent since the Act was passed – even as emissions of the six main pollutants it targets have fallen by 71 percent. Despite these gains, not everyone is seeing cleaner skies. The ALA found that 18 million Americans "live in 12 counties with unhealthful levels of all three [pollutants tracked in the study]: ozone and short-term and year-round particle pollution." And even as ozone and year-round particle pollution have dropped, it observed "an unrelenting increase in dangerous spikes in particle pollution." For both short-term "spikes" and a year-round presence of dust, smoke, and soot particles, California cities topped the rankings, and some, like Bakersfield in the Central Valley, indeed saw an increase in this type of pollution. The Golden State's hill-and-valley topography helps create "thermal inversions," layers of warm air that trap smog in low-lying areas. While cities like Los Angeles have been grappling with this problem for decades, Professor Anastasio says that recent environmental changes may have driven the latest increase. "My guess is that the California drought is partially responsible for making fine particle ... pollution worse since 2010," he writes in an email. "Winter rains clean the air of particles, but we had little precipitation for the past five years, until this year." Professor Cappa, his colleague at UC-Davis's Air Quality Research Center, adds that the drought has brought an increase of wildfires. "We've had a lot of pretty bad fire seasons lately, which can really lead to poor air quality," he says. Cappa cautions against making too much of the recent increase. "We want to pay attention to those [spikes]," he explains, "but we don't necessarily want to overly worry right away, because that ... probably is linked to short-term changes in meteorology and climate in the area." Both droughts and wildfires, he points out, are linked to climate change, meaning that curbing them could require stemming CO2 emissions. But at the same time, environmental advocate Nayamin Martinez also sees a need for stronger regulations at the local level. Ms. Martinez serves as director of the Central California Environmental Justice Network (CCEJN); in her group's view, the San Joaquin Valley Air Pollution Control District, which oversees emissions-reduction plans in Bakersfield, Fresno, and some of the other most-polluted cities, "could be coming up with regulations that are more stringent that could protect our air." In a phone interview with the Monitor, Martinez points to one recent success in this effort: Advocacy from CCEJN and other environmental groups convinced California's Air Resources Board to reject San Joaquin's plan to reduce particulate pollution, requiring it to come up with a more aggressive one. In the nation's remaining pollution pockets, the ALA's research bolsters the case for improvements like these. When Martinez spoke with the Monitor on Wednesday afternoon, the local ALA chapter had already briefed her on the latest report. "Having this type of national organization coming out with this information really backs up our argument," she says.

Figure 2. Clustering of Sentences. Highlights with the same color represent a cluster.

Title: New report finds cleaner air for many, but not all.

More Americans have been breathing easy in recent years. On Wednesday, the American Lung Association (ALA) released its annual "State of the Air" report, which tracks air quality for the years 2013-2015, both nationally and for specific metrics, like cities and types of pollution. While some areas showed continued improvement and others deterioration, overall the report found that "the number of people exposed to unhealthy levels of air pollution dropped to more than 125 million people," down almost 25 percent from 166 million in the years covered in the previous report (2012 – 2014). The report credits this progress to the Clean Air Act, one of the US environmental movement's landmark achievements passed in 1970, and the scientists who study this issue agree. But in some parts of the country – especially California – environmental advocates see a need for smart policy decisions that will continue cleaning the air. "This [report] is just further evidence that our efforts to clean up the air are working, but that we have to push further to help improve the air ... for everybody in the country," Chris Cappa, a professor in the department of civil and environmental engineering at University of California–Davis, tells The Christian Science Monitor in a phone interview. Air pollution in such places as Beijing and New Delhi has made headlines in recent years, but the mid-20th century saw many American cities wrapped in a similar haze. In October 1948, the air over Denora, Penn., got so thick with smog from its steel and wire factory that the industrial town's fire department handed out oxygen tanks to ailing residents. The pollution killed 20 people and left half of the town's 14,000 residents recovering in hospitals or at home. A series of policies meant to prevent this disaster from happening again culminated in 1970 with the Clean Air Act's passage. **Under the law, the US Environmental Protection Agency (EPA) must set science-based air quality standards for common types of air quality pollutants, and requires states to set enforceable limits on meeting those standards**. Other components of the act cover interstate pollution and mobile emitters, such as car tailpipes. "While there are other factors" behind the drop, "the Clean Air Act has probably been the most important component," explains Cort Anastasio, professor in the air, land, and water resources department at UC-Davis, in an email to the Monitor. "EPA regulations," he continues, "while regularly derided by some in (and out) of politics, are making our air cleaner and provide benefits (in the form of increased human health) that far outweigh their costs." **Contrary to concerns that these regulations would stifle businesses, the EPA estimates that just one provision of the act – 1990 amendments that target acid rain – will yield an economic benefit of $2 trillion by 2020, at an enforcement cost of $65 billion**. Already, the ALA's report notes that America's gross domestic product has grown by 246 percent since the Act was passed – even as emissions of the six main pollutants it targets have fallen by 71 percent. Despite these gains, not everyone is seeing cleaner skies. The ALA found that 18 million Americans "live in 12 counties with unhealthful levels of all three [pollutants tracked in the study]: ozone and short-term and year-round particle pollution." And even as ozone and year-round particle pollution have dropped, it observed "an unrelenting increase in dangerous spikes in particle pollution." For both short-term "spikes" and a year-round presence of dust, smoke, and soot particles, California cities topped the rankings, and some, like Bakersfield in the Central Valley, indeed saw an increase in this type of pollution. The Golden State's hill-and-valley topography helps create "thermal inversions," layers of warm air that trap smog in low-lying areas. While cities like Los Angeles have been grappling with this problem for decades, Professor Anastasio says that recent environmental changes may have driven the latest increase. "My guess is that the California drought is partially responsible for making fine particle ... pollution worse since 2010," he writes in an email. "Winter rains clean the air of particles, but we had little precipitation for the past five years, until this year." Professor Cappa, his colleague at UC-Davis's Air Quality Research Center, adds that the drought has brought an increase of wildfires. "We've had a lot of pretty bad fire seasons lately, which can really lead to poor air quality," he says. Cappa cautions against making too much of the recent increase. **"We want to pay attention to those [spikes]," he explains, "but we don't necessarily want to overly worry right away, because that ... probably is linked to short-term changes in meteorology and climate in the area."** Both droughts and wildfires, he points out, are linked to climate change, meaning that curbing them could require stemming CO_2 emissions. But at the same time, environmental advocate Nayamin Martinez also sees a need for stronger regulations at the local level. Ms. Martinez serves as director of the Central California Environmental Justice Network (CCEJN); in her group's view, the San Joaquin Valley Air Pollution Control District, which oversees emissions-reduction plans in Bakersfield, Fresno, and some of the other most-polluted cities, "could be coming up with regulations that are more stringent that could protect our air." In a phone interview with the Monitor, Martinez points to one recent success in this effort: Advocacy from CCEJN and other environmental groups convinced California's Air Resources Board to reject San Joaquin's plan to reduce particulate pollution, requiring it to come up with a more aggressive one. In the nation's remaining pollution pockets, the ALA's research bolsters the case for improvements like these. When Martinez spoke with the Monitor on Wednesday afternoon, the local ALA chapter had already briefed her on the latest report. **"Having this type of national organization coming out with this information really backs up our argument," she says.**

Figure 3. Summary 1 (TF-ISF measure)

Title: New report finds cleaner air for many, but not all.

More Americans have been breathing easy in recent years. On Wednesday, the American Lung Association (ALA) released its annual "State of the Air" report, which tracks air quality for the years 2013-2015, both nationally and for specific metrics, like cities and types of pollution. While some areas showed continued improvement and others deterioration, overall the report found that "the number of people exposed to unhealthy levels of air pollution dropped to more than 125 million people," down almost 25 percent from 166 million in the years covered in the previous report (2012 – 2014). The report credits this progress to the Clean Air Act, one of the US environmental movement's landmark achievements passed in 1970, and the scientists who study this issue agree. **But in some parts of the country – especially California – environmental advocates see a need for smart policy decisions that will continue cleaning the air**. "This [report] is just further evidence that our efforts to clean up the air are working, but that we have to push further to help improve the air ... for everybody in the country," Chris Cappa, a professor in the department of civil and environmental engineering at University of California–Davis, tells The Christian Science Monitor in a phone interview. Air pollution in such places as Beijing and New Delhi has made headlines in recent years, but the mid-20th century saw many American cities wrapped in a similar haze. In October 1948, the air over Denora, Penn., got so thick with smog from its steel and wire factory that the industrial town's fire department handed out oxygen tanks to ailing residents. The pollution killed 20 people and left half of the town's 14,000 residents recovering in hospitals or at home. A series of policies meant to prevent this disaster from happening again culminated in 1970 with the Clean Air Act's passage. Under the law, the US Environmental Protection Agency (EPA) must set science-based air quality standards for common types of air quality pollutants, and requires states to set enforceable limits on meeting those standards. Other components of the act cover interstate pollution and mobile emitters, such as car tailpipes. "While there are other factors" behind the drop, "the Clean Air Act has probably been the most important component," explains Cort Anastasio, professor in the air, land, and water resources department at UC-Davis, in an email to the Monitor. "EPA regulations," he continues, "while regularly derided by some in (and out) of politics, are making our air cleaner and provide benefits (in the form of increased human health) that far outweigh their costs." Contrary to concerns that these regulations would stifle businesses, the EPA estimates that just one provision of the act – 1990 amendments that target acid rain – will yield an economic benefit of $2 trillion by 2020, at an enforcement cost of $65 billion. **Already, the ALA's report notes that America's gross domestic product has grown by 246 percent since the Act was passed – even as emissions of the six main pollutants it targets have fallen by 71 percent**. Despite these gains, not everyone is seeing cleaner skies. The ALA found that 18 million Americans "live in 12 counties with unhealthful levels of all three [pollutants tracked in the study]: ozone and short-term and year-round particle pollution." And even as ozone and year-round particle pollution have dropped, it observed "an unrelenting increase in dangerous spikes in particle pollution." For both short-term "spikes" and a year-round presence of dust, smoke, and soot particles, California cities topped the rankings, and some, like Bakersfield in the Central Valley, indeed saw an increase in this type of pollution. The Golden State's hill-and-valley topography helps create "thermal inversions," layers of warm air that trap smog in low-lying areas. While cities like Los Angeles have been grappling with this problem for decades, Professor Anastasio says that recent environmental changes may have driven the latest increase. "My guess is that the California drought is partially responsible for making fine particle ... pollution worse since 2010," he writes in an email. "Winter rains clean the air of particles, but we had little precipitation for the past five years, until this year." Professor Cappa, his colleague at UC-Davis's Air Quality Research Center, adds that the drought has brought an increase of wildfires. "We've had a lot of pretty bad fire seasons lately, which can really lead to poor air quality," he says. Cappa cautions against making too much of the recent increase. **"We want to pay attention to those [spikes]," he explains, "but we don't necessarily want to overly worry right away, because that ... probably is linked to short-term changes in meteorology and climate in the area."** Both droughts and wildfires, he points out, are linked to climate change, meaning that curbing them could require stemming CO2 emissions. But at the same time, environmental advocate Nayamin Martinez also sees a need for stronger regulations at the local level. Ms. Martinez serves as director of the Central California Environmental Justice Network (CCEJN); in her group's view, the San Joaquin Valley Air Pollution Control District, which oversees emissions-reduction plans in Bakersfield, Fresno, and some of the other most-polluted cities, "could be coming up with regulations that are more stringent that could protect our air." In a phone interview with the Monitor, Martinez points to one recent success in this effort: Advocacy from CCEJN and other environmental groups convinced California's Air Resources Board to reject San Joaquin's plan to reduce particulate pollution, requiring it to come up with a more aggressive one. In the nation's remaining pollution pockets, the ALA's research bolsters the case for improvements like these. When Martinez spoke with the Monitor on Wednesday afternoon, the local ALA chapter had already briefed her on the latest report. **"Having this type of national organization coming out with this information really backs up our argument," she says.**

Figure 4. Summary 2 (Headline to sentence similarity)