

# LING 573: Document Summarization Project Report

Daniel Campos, Sicong Huang, Shunjie Wang, Simola Nayak, and Hayley Luke

University of Washington

{dacamp, huangs33, shunjiew, simnayak, jhluke}@uw.edu

## Abstract

We design and implement Mockingbird, a topic-focused multi-document extractive summarization system. Building on the LexRank graph algorithm our system uses sentence similarity and topic-sentence bias to produce candidates. Next, the ranked sentences are selected to limit redundancy and stay under 100 words. Our system outperforms the LEAD and MEAD baseline by a fair margin. Future work will focus on forms of text representation and processing along with more complex selection and sorting can improve system performance

## 1 Introduction

Topic oriented document clusters like ACQUAINT (Graff, 2002) and ACQUAINT-2 have been used as a starting point to explore various methods for document summarization. More specifically, these corpuses have been used to study extractive multi-document topic summarization. These corpuses have been the focus of study in TAC(Text Analytics Conference) (Dang and Owczarzak, 2008) summarization shared task. In the formalization of this task give a topic and a set of news wire documents a competitive system should create a high quality summary of the topic using sentences from the documents. Systems are expected to produce summaries up to 100 words and summaries should be coherent and not contain duplication. Once summaries are generated for all the topics being studied, methods are evaluated and compared using the standard ROUGE metric (Lin, 2004).

Our exploratory system, called Mockingbird(MB), is based on the Biased LexRank graph approach (Otterbacher et al., 2009). In this approach, we provide a ranking of all candidate sentences by combining a matrix which represents sentence similarity with a topic and sentence similarity. After a ranking is produced sentences are selected to maximize their lexrank score but minimize duplication of content. Our method relies on word vectors

(Mikolov et al., 2013) from the spacy library<sup>1</sup> to represent sentences and we experiment with the effects of evaluating complete sentences, sentences without stop words, and only the nouns in the sentences. After we have a set of candidate sentences we sort them in reverse chronological order and then realize the content to match the TAC output format. To understand our system performance we compare our systems ROUGE-1 and ROUGE-2 scores when compared to LEAD, and MEAD baselines. Our system across the board performs favorably as we beat the two baselines(MEAD and LEAD). However, our system lags behind the top system from TAC 2010 by a significant margin.

## 2 System Overview

Mockingbird has been designed as a simple end-to-end pipeline with the goal of producing a structure we can continue to tweak and modify to understand the effect of various changes have on downstream performance. The pipeline, represented in 1, broadly has 4 steps: document input and processing, content selection, information ordering, content realization.

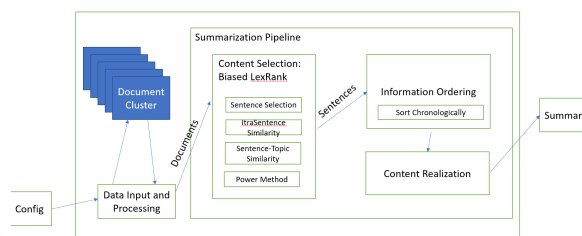


Figure 1: Overview Of Mockingbirds Architecture

<sup>1</sup><https://spacy.io/>

### 3 Approach

In this section we will describe the in detail each of the steps our system takes to summarize topics.

#### 3.1 Data Input and Processing

Documents come from the ACQUAINT and ACQUAINT-2 corpus and are a mixture of HTML and XML. The document pipeline takes as an input a configuration file(XML) which details a series of topics and associates the them with a group of document ID's. Using this configuration file, we compile a list of topics, and use the document ID's to determine the path to the relevant corpus file on disk. We then search the file for the information relevant to our specific document ID, and extract the text, the date, and the title. We clean the data, converting the date to a usable format, and stripping excess white space and symbols in the text. Finally, we naively split the text on specific punctuation marks to retrieve a list of sentences.

#### 3.2 Content Selection

Next, there is the content selection pipeline. In this step we consume the structured information from the processing pipeline and produce a candidate set of sentences. Our system selects content using a modified version of the Biased LexRank Graph approach. This method computes a intrasentence and topic-sentence similarity which is used to create a similarity matrix and a bias vector. In this method, each node of the graph represents a sentence and edges between nodes are represented by their intrasentence similarity. This representation allows ranking of sentences to be based on their similarity to the topic and their centrality in the topic document clustering. Our method differs from the original implementation of Lexrank(Otterbacher et al., 2005) as we leverage word embeddings instead of their tf-idf implementation.

For each topic we first assemble all sentences in scope of the topic that have at least 4 words. Then, for each sentence, we use spacy to create a vector representation for each sentence. This method leverages the word embeddings known as GloVe (Pennington et al., 2014) and produces a sentence embedding by averaging the word vectors of each word in the sentence. We experiment with various text normalization methods like stop-word removal and nouns only but broadly the MB pipeline is agnostic. We follow equation 1 to we generate a two dimension matrix representing the cosine similar-

ity between all the sentences related to the topic. Additionally, we create a bias 1d vector which represents the cosine similarity between the topic and each sentence. For inter-sentential cosine similarity, we use a minimum threshold of 0.3 as we find word similarity tends to be higher then tf-idf similarity.

$$similarity(u, v) = \cos(u, v) = \frac{u * v}{\|u\| \|v\|} \quad (1)$$

Using the bias vector and the inter sentence similarity matrix we compute a lexrak for each sentence using the equation 4 where  $s$  is a sentence,  $q$  is the topic string,  $d$  is the weight bias for the topic(set to 0.8), and  $S$  represents all sentence in the topic.

$$bias(s|q) = \frac{\cos(s, q)}{\sum_{a \in S} \cos(a, q)} \quad (2)$$

$$sentsim(s, S) = \sum_{s_1 \in S} \frac{\cos(s, s_1)}{\sum_{s_2 \in S} \cos(s_1, s_2)} \quad (3)$$

$$lr(s|q) = d * bias(s|q) + (1 - d) * sentsim(s, S) \quad (4)$$

In our system, we turn our 2d inter-sentential similarity matrix( $IS$ ) and bias vector( $BV$ ) into a matrix  $M$  using equation 5. Then, we implement the power method, as described in equation 6, to produce a converged LexRank values for each sentence. As LexRank represents sentence probabilities we initial have a uniform distribution (all sentences are equally likely to show up) and we use the power method until the method converges or an  $\epsilon$  of 0.3.

$$M = [d * IS + (1 - d) * BV]^T \quad (5)$$

$$P_t = M * p_t - 1 \quad (6)$$

Once we have assembled sampling probabilities for each sentence we select sentences until we have either used all candidate sentences or reached the target length of 100 words. Sentences are initially ranked by LexRank score and we select candidates that have less than a 0.6 cosine similarity to all other selected sentence. This filter is applied to minimize sentence redundancy.

### 3.3 Information Ordering

Following content selection we run our information ordering system. We used a basic chronological ordering approach, by sorting our sentence objects by their `doc_date` property.

### 3.4 Content Realization

Following Information ordering we run our content realization system. Currently, this system only formats the text to match the desired file format.

### 3.5 Evaluation

Once all summaries have been created our system moves onto evaluation.

## 4 Results

To evaluate our system performance we ran our system on the 2010 TAC shared evaluation task. The 2010 TAC task has 43 systems including baselines. The 1st baseline (LEAD), created summaries using the first 100 words from the most recent document. The 2nd baseline was the output of the MEAD summarization system (Radev et al., 2003). We have also included system 43 and system 22 as benchmarks because they had the best performance in the shared task (eval and dev set respectively). We evaluate models according to their ROUGE-2 and ROUGE-1 Recall scores on the devtest portion of the 2010 TAC task. Our system, has three variants indicating various methods of sentence representation. Default uses the sentence average word embeddings without normalization. No Stop words uses sentence average embedding with stop word removed. Nouns uses sentence average embeddings only based on the nouns in the text.

Looking at ??, we can see that our system, Mockingbird outperforms the baselines by a fair margin. Our system runs best when removing stop words as this shows a 0.02324 improvement over the default system. While our model is competitive with baselines it is outperformed by a fair margin by system 22, which had the best performance on the devtest set for 2010 TAC.

## 5 Discussion

We were surprised how well Mockingbird performed given the simplicity of the system. Given our system design was focused on producing a functional end to end summarization pipeline we were not expecting to beat the MEAD baseline. Given

System Name	ROUGE-2	ROUGE-1
LEAD	0.05376	-
MEAD	0.05927	-
MB(Default)	0.05047	0.23737
MB(Nouns)	0.06704	0.32014
MB(No Stop words)	0.07371	0.33193
System 43	0.01154	-
System 22	<b>0.09574</b>	-

the focus on a functional platform we performed no hyperparameter tuning for all the static values of our lexicrank models and thus believe we could improve our system further once we find the optimal values. Moreover, since most of our hyperparameters were based on prior literature and those methodologies rely on tf-idf methodologies we expect our optimal parameters will likely be very different than our current settings. b We were also pleasantly surprised to find a huge improvement in performance by performing some basic sentence normalization for our sentence representation for vectors. This leads us to believe that there are likely many other sentence normalization approaches that could lead to further gains. Outside of the content selection model improvements we believe there are likely further gains to be had by implementing less naive strategies in information ordering and content realization.

## 6 Conclusion and Future Work

Mocking bird(MB) is a topic-based multi-document extractive text summarization that leverages word vectors to select salient sentences. MB is a end to end summarization system that has been implemented in a simple and expandable fashion with the goal of allowing quick tweaks for exploration of model performance.

MB uses a modified version of the LexRank similarity graph method which uses inter-sentence similarity and topic sentence similarity to produce a sampling probability for each sentence in a topic. Then, MB select greedily from ranked candidate sentences to reach the 100 word summarization target producing relevant summaries that have little redundancy.

MB outperforms the LEAD and MEAD baselines for the 2010 TAC summarization track on the ROUGE-2 scores on the devtest set. MB does best when sentences are pre-process to remove stop words but there is likely substantially more fine

tuning that can be done to further improve model scores. MB has had no hyperparameter tuning and thus is likely not performing at its global maxima. In future iterations of MB we will tune our hyperparameters, explore other sentence representations like BERT (Devlin et al., 2019), implement more text normalization, and implement more complex and complete information ordering and content realization systems.

## References

- Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. *Theory and Applications of Categories*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.
- David Steven Graff. 2002. The acquaint corpus of english news text.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.
- Jahna Otterbacher, Güneş Erkan, and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 915–922, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Jahna Otterbacher, Güneş Erkan, and Dragomir R. Radev. 2009. Biased lexicrank: Passage retrieval using random walks with question-based priors. *Inf. Process. Manag.*, 45:42–54.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.
- Dragomir R. Radev, Jahna Otterbacher, Hong Qi, and Daniel See Wai Tam. 2003. Mead reduces: Michigan at duc 2003.