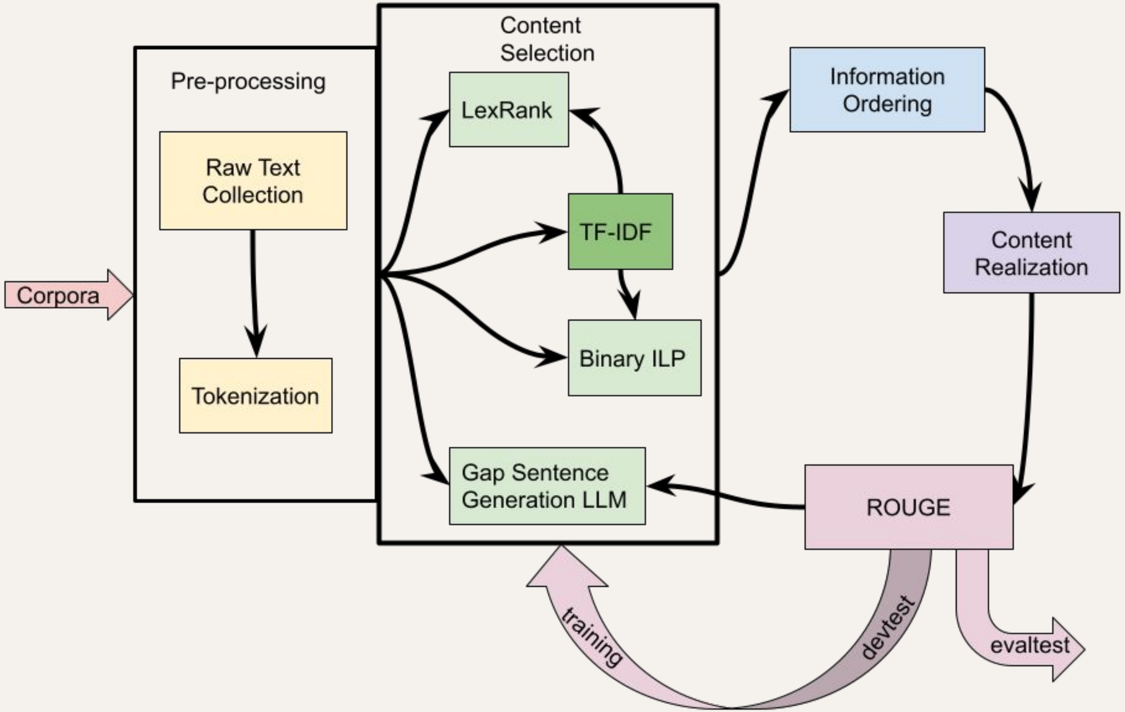
D3
The Initial System

Anna Batra, Sam Briggs, Junyin Chen, Hilly Steinmetz

Table of Contents

System Architecture



TF-IDF

$$tf \cdot idf(t, d, D) = tf(t, d) \cdot idf(t, d, D)$$

TF Log Normalization

$$\begin{aligned} tf(t, d) &= \log(\delta_1 + f_{t,d}) \\ &= \log(\delta_1 + |\{t \mid t \in d, d \in D\}|) \end{aligned}$$

Smoothed IDF

$$\begin{aligned} idf(t, D) &= \delta_2 + \log\left(\frac{N}{\delta_2 + n_t}\right) \\ &= \delta_2 + \log\left(\frac{|D|}{\delta_2 + |\{d \mid t \in d, d \in D\}|}\right) \end{aligned}$$

TF-IDF Summarizer

Max

Max unigram TF-IDF weight over a sentence

Average

Averaged TF-IDF weight over unigrams in sentence

TODO: Ranking

- Pick top N sentences
- Keep sentence ordering for summary

ILP - Summarization

$$\begin{array}{ll} \text{maxmimize} & \sum_i w_i z_i \\ \text{Subject To} & \sum_j A_{i,j} y_j \geq z_i \\ & A_{i,j} \leq z_i \\ & \sum_j l_j y_j \geq L \end{array}$$

ILP - What we did

pulp

CBC MILP Solver
v. 2.10.3

Concepts

Unigrams

Weights

TF-IDF of the
unigram

Sentence Ordering

Same order as
docset/doc system

LexRank – Approach

1) Concatenate Documents

Adapt LexRank for multidoc summarization

2) Create TF-IDF dictionary

Collect term frequency and inverse doc frequencies

3) Generate Sentence Graph

Create similarity matrix with modified cosine similarity

4) Rank Sentences by Importance

Use the power method to find eigenvalue of matrix

LexRank – Equations

$$\text{idf-modified-cosine}(x, y) = \frac{\sum_{w \in x,y} \text{tf}_{w,x}\text{tf}_{w,y}(\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x}(\text{tf}_{x_i,x}\text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(\text{tf}_{y_i,y}\text{idf}_{y_i})^2}}$$

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{p(v)}{deg(v)}$$

$$\mathbf{p} = [d\mathbf{U} + (1-d)\mathbf{B}]^{\mathrm{T}}\mathbf{p}$$

LexRank – Implementation choices

- Didn't implement algorithm paper directly – changed the regularization term when setting up the matrix

- This change ensures that the matrix satisfies the property of a stochastic matrix that all rows sum to 1

MInputAn array S of n sentences, cosine threshold t **output**: An array L of LexRank scores
Array $CosineMatrix[n][n]$;
Array $Degree[n]$;
Array $L[n]$;
for $i \leftarrow 1$ *to* n **do**
    for $j \leftarrow 1$ *to* n **do**
        $CosineMatrix[i][j] =$ `idf-modified-cosine`$(S[i], S[j])$;
        if $CosineMatrix[i][j] > t$ **then**
            $CosineMatrix[i][j] = 1$;
            ~~$Degree[i]++$;~~
        end
        else
            $CosineMatrix[i][j] = 0$;
        end
    end
end
for $i \leftarrow 1$ *to* n **do**
    for $j \leftarrow 1$ *to* n **do**
        $CosineMatrix[i][j] =$ ~~$CosineMatrix[i][j]/Degree[i]$;~~
    end
end
$L =$ `PowerMethod`$(CosineMatrix, n, \epsilon)$;
return L;

CosineMatrix[*i*][*j*] / ∑ {1 → *n*) CosineMatrix[*i*][*j*]

LexRank – Easy with NumPy!

```
def power_method(
        matrix: np.ndarray,
        error: float,
        d: float
    ) -> np.ndarray:
    '''
    Power method for solving stochastic, irreducible, aperiodic matrices
    Arguments:
        - matrix: a square matrix
        - error: when the error is low enough to finish algorithm
        - d: dampening factor (to ensure convergence)
    '''
    p_t = np.ones(shape=(matrix.shape[0]))/matrix.shape[0]
    t, delta = 0, None
    U = np.ones(shape=matrix.shape)/matrix.shape[0]
    while delta is None or delta > error:
        t += 1
        p_t_1 = p_t
        p_t = np.matmul(
            (U * d) + (matrix.T * (1-d)),
            p_t
        )
        delta = np.linalg.norm(p_t - p_t_1)
    # normalize ranking
    p_t = p_t/p_t.sum()
    return p_t
```

LLM

- Rank sentence based on Gap Sentences Generation introduced in Zhang et al. (2019) for training Pegasus
  - Select top m sentences based on the ROUGE-1's F1 score between the selected sentence and the rest of the document
  - Discard lower 50% of the sentence based on the ROUGE score to truncate input sequence to 1024 token
- Use "google/pegasus-cnn_dailymail" model for training with batch size of 6 and epoch of 12
  - "google/T5-small" model does not produce complete sentences
  - "google/pegasus-xsum" raise CUDA not initialized error.

Algorithm 1 Independent sentence selection

$D := \{x_i\}_n \leftarrow$ sentences in *document*
$S := \emptyset$
$I \leftarrow$ list contains index from 0 to n
for $j \leftarrow 1$ to n **do**
    $s_i :=$ *rouge*$(x_i, D \setminus \{x_i\})$
    $S := S \cup \{s_i\}$
$I :=$ *sort*(I) Based on the value in S

Results

	ROUGE1	ROUGE2
Binary ILP	0.12085	0.01533
LexRank	0.13720	0.02341
GSG LLM	0.21037	0.06214

ROUGE Recall Scores

Error Analysis

gold	In the worst school killing in U.S. history, two students at Columbine High School in Littleton, Colorado, a Denver suburb, entered their school on Tuesday, April 20, 1999, to shoot and bomb. At the end 15 were dead and dozens injured. The dead included the two students, Eric Harris and Dylan Klebold, who killed themselves. Harris and Klebold were enraged by what they considered taunts and insults from classmates and had planned the massacre for more than a year. The school is a sealed crime scene and Columbine students will complete the school year at a nearby high school.
Binary ILP	At one point , two bomb squad trucks sped to the school after a backpack scare . Phone : (888) 603-1036 Please comfort this town . '' Many looked for it Saturday morning on top of Mt . But what community was it from ? There are the communities that existed already , like Columbine students and Columbine Valley residents . Brothers Jonathan and Stephen Cohen sang a tribute they wrote . `` Columbine ! '' `` Love is stronger than death . '' Some players said the donations and support will encourage them to play better .
LexRank	Sheriff John Stone said Tuesday afternoon that there could be as many as 25 dead. The arrival of two bomb squad trucks with sirens blaring further shook those inside. With photo. With photo. The New York Times plans two pages of stories, photos and graphics on the aftermath of the school shooting in a Denver suburb that left 15 dead. Herbert interjected:` `I'm a little worried about putting all the kids in one place. Littleton needs comfort. Littleton needs comfort.
GSG LLM	Sheriff's initial estimate of as many as 25 dead in Columbine massacre was off the mark. discrepancy occurred because the SWAT teams that picked their way past bombs and bodies in an effort to secure building covered overlapping areas.

Discussion (Issues & Successes/ Future Improvements)

- Finish information ordering and content realization for TF-IDF
- Use dev-test to pick good delta for smoothing in TF-IDF (TF-IDF & ILP Summarization)
- Detokenizing summary for ILP
- Add bigram concepts to ILP, dev-test to see if better than unigram
- Switch to BLOOM with AutoModelForQuestionAnswering class for training, since using AutoModelForCausalLM for BLOOM produce input size not match error.
- Improve LLM pre-processing methods

Packages

- Pulp
- Transformer
- NLTK
- Datasets
- Evaluate
- Rouge-score
- Pytorch

Challenges

- Had to handle exceptional data in document set
- Algorithm described in papers didn’t always line up with description in papers
- Condor couldn’t handle data without decreasing batch sizes to very low

References

- Yohei Seki. 2003. Sentence extraction by tf/idf and position weighting from newspaper articles. National Institute of Informatics: Proceedings of the Third NTCIR Workshop.
- Luo, W., Liu, F., Liu, Z., & Litman, D. (2018). A novel ILP framework for summarizing content with high lexical variety. *Natural Language Engineering, 24*(6), 887-920. doi:10.1017/S1351324918000323
- Erkan, Günes, and Dragomir R. Radev. 2004. "Lexrank: Graph-based lexical centrality as salience in text summarization." in *Journal of artificial intelligence research* 22: 457-479.
- Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." International Conference on Machine Learning. PMLR, 2020.

Live Demo