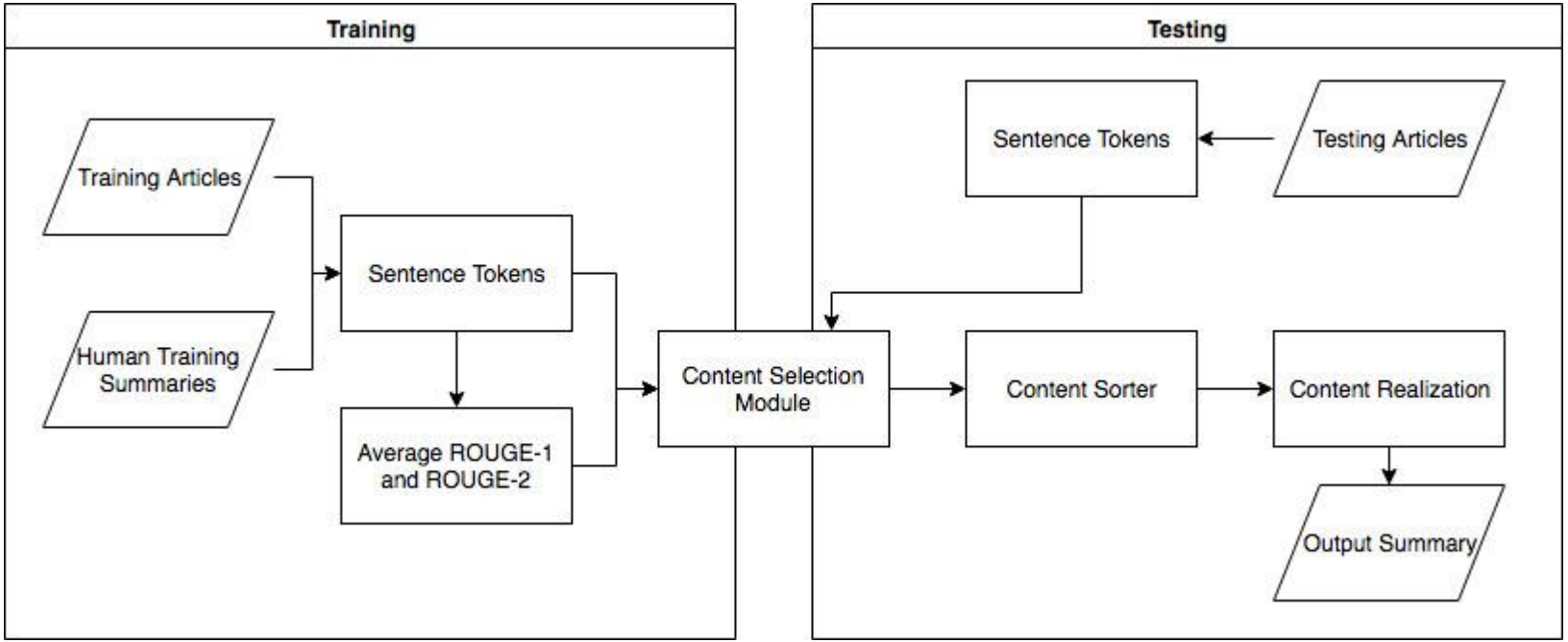
573 Project Report - D3

Mackie Blackburn, Xi Chen, Yuan Zhang

System Overview



Improvements in Preprocessing

Streamlined preprocessing:

Integrated preprocessing with data extraction and preparation.

Preprocessing steps:

sentence → lowercased, stop-worded, lemmatized (n. & v.),
non-alphanumeric characters removed → list of word tokens

Cached two parallel dictionaries: one with processed sent.s and the other with original sent.s for easy lookup

Topic Orientation

Adopted query-based LexRank approach (Erkan and Radev, 2005)

Combined relevance score (sent to topic) and salience score (sent to sent)

$$p(s|q) = d\frac{\text{rel}(s|q)}{\sum_{z\in C}\text{rel}(z|q)} + (1-d)\sum_{v\in C}\frac{\text{sim}(s,v)}{\sum_{z\in C} sim(z,v)}p(v|q)$$

Markov Random Walk: power method to get eigenvector for convergence

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

Data: Removed SummBank data (no topics); Added DUC 2007 data

Improvements in Content Selection

Added Features

- Lexrank
- Query-Based Lexrank
- Sentence index, first sentences

Fixed math bug in LLR

Feature	Score
Query-Based Lexrank	0.133033
LLR	0.121160
Sentence Length	0.080823
Earliest First Occurrences	0.065079
LLR Sum	0.064063
Count of NN	0.056965
Is First Sentence	0.054739
Sentence Index	0.039904
Average First Occurrences	0.037592
TF*IDF Average	0.032323
Reverse KL Divergence of Bigrams	0.028661
Lexrank	0.027235
KL Divergence of Bigrams	0.025725
KL Divergence of Unigrams	0.025337
Average Position of words in Documents	0.025042
TF*IDF Sum	0.019876
Count of JJ	0.019386
Reverse KL Divergence of Unigrams	0.01817
Sentiment Intensity Score	0.017199
Probability of Number	0.014041
Count of VBP	0.012650
Count of VB	0.011904

Information Ordering

Due to sparsity of training data, we apply a semi-supervised algorithm to order sentences picked up by the content selector. The algorithm is based on the paper ‘Sentence Ordering based Cluster Adjacency in Multi-Document Summarization’ by DongHong and Yu (2008).

Information Ordering

Basic Idea of the algorithm:

Suppose we have the co-occurrence probability $CO_{m,n}$,between each sentence pair in the summary $\{S_1, S_2, \ldots, S_{len(summary)}\}$.

If we know the *k*th sentence in the summary is S_i, then we can always choose the (*k+1*)th sentence by choosing the one with maximum $CO_{i,j}$.

However, the co-occurrence probability $CO_{m,n}$ is practically always zero...

Information Ordering

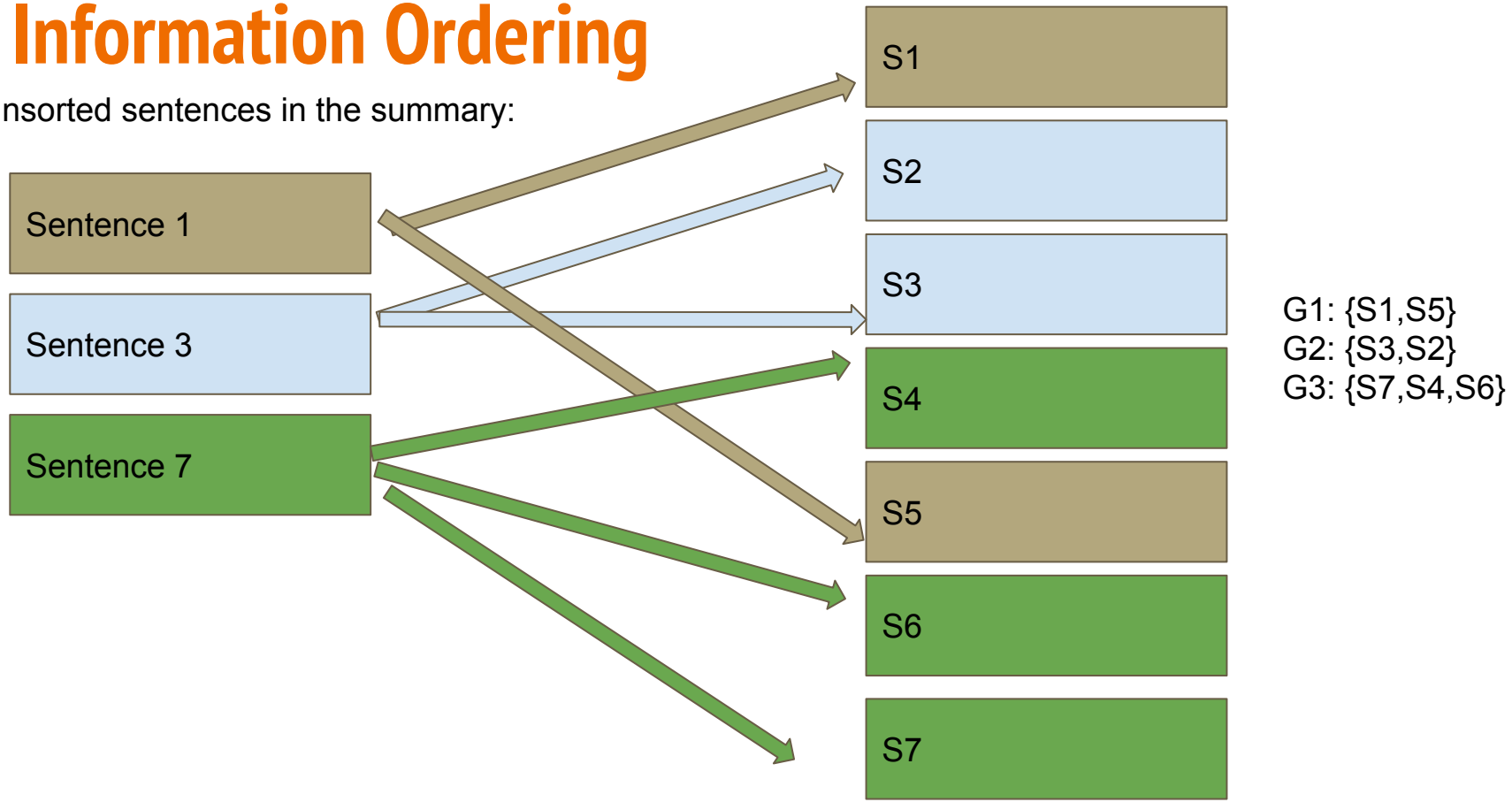
As the result, we augment each sentence in the summary into a sentence group by clustering. Then we approximate sentence co-occurrence $CO_{m,n}$ by sentence group co-occurrence probability:

$$C_{m,n} = f(G_m, G_n)^2 / (f(G_m)f(G_n))$$

Here the $f(G_m, G_n)$ is the sentence group co-occurrence frequency within a word window and $f(G_m)$ is the sentence group co-occurrence frequency. This probability is about sentence groups' adjacency to each other.

Information Ordering

Unsorted sentences in the summary:



Information Ordering(*)

Implementation:

[1]Use glove 50D word embedding to convert each sentence into vector

[2]Based on the vectors, run label spreading clustering to get groups

[3]Calculate group based co-occurrence probabilities

[4]Run greedy picking up based on $C_{m,n}$

Information Ordering

Evaluation:

The evaluation metric of an ordering is Kendall's τ:

$$\tau = 1 - 2(\text{numbers\_of\_inverions}) / (N(N-1)/2)$$

Kendall's τ is always between (-1, 1). τ of -1 means a totally reversed order, τ of 1 means totally ordered, and τ of 0 means the ordering is random.

Information Ordering

Evaluation Dataset: 20 human extracted passages (of 3~4 sentences each) from training data, evaluate on algorithm output vs human summaries.

Model name:	τ
Random:	0
Adjacency $_{\text{(symmetric window size = 2)}}$:	0.200
Adjacency $_{\text{(symmetric window size = 1)}}$:	0.324
Adjacency $_{\text{(forward window size = 1)}}$:	0.356
Chronological:	0.465

Score Improvement

Average Recall Results on Devtest Data

Model	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
D2	0.18687	0.04579	0.01503	0.00558
D3	0.22459	0.05354	0.01287	0.00424

Issues and Successes

Topic-Focused Lexrank is a very good feature

Adding topic focus doesn’t always improve ROUGE

- KL divergence of sentence from topic
- Topic focused features may favor sentences with similar information

Summary Examples

The British government set targets on obesity because it increases the likelihood of coronary heart disease, strokes and illnesses including diabetes. Over 12 percent said they did not eat breakfast, and close to 30 percent were unsatisfied with their weight. Several factors contribute to the higher prevalence of obesity in adult women, Al-Awadi said. Kuwaiti women accounted for 50.4 percent of the country's population, which is 708,000. Fifteen percent of female adults suffer from obesity, while the level among male adults 10.68 percent. The ratio of boys is 14.7 percent, almost double that of girls. According to his study, 42 percent of Kuwaiti women and 28 percent of men are obese.

Planned Improvements

Larger background corpus for LLR

- New York Times on Patas

Try extra features in similarity calculation, such as publish date(?)

- Find more paper related

Find a better way to pick the first sentence.

References

[1] G. Erkan and D. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artificial Intelligence Research, 22(1):457-479*, 2004.

[2] J. Otterbacher G. Erkan and D. R. Radev. Using random walks for question-focused sentence retrieval. *Proceedings of Human Languages Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp.915-922*, 2005.

[3] K. Hong and A. Nenkova. Improving the estimation of word importance for news multi- document summarization. *in Proceedings of EACL*, 2014.

[4] P.E. Genest G. Lapalme and M. Yousfi-Monod. Hextac: The creation of a manual extractive run. *Proceedings of the Second Text Analysis Conference (TAC 2009)*, 2009.

[5] A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 2011.