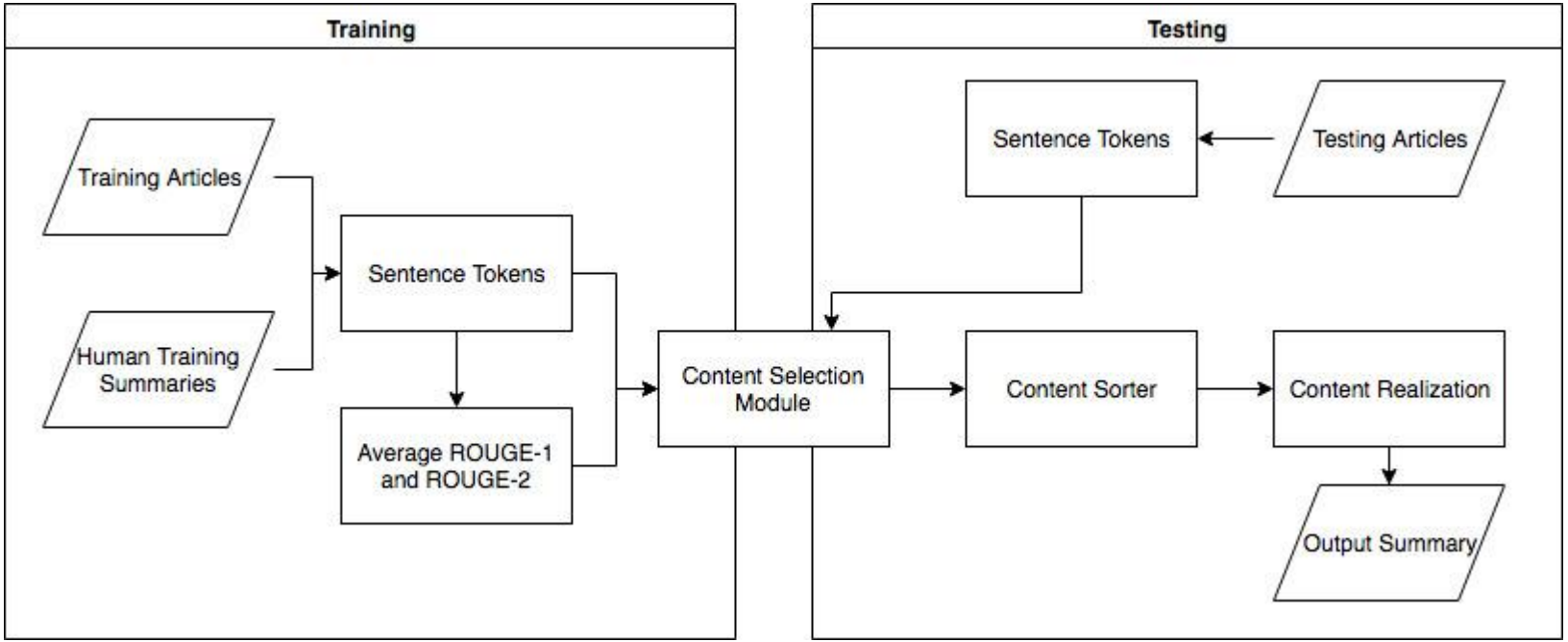
573 Project Report - D4

Mackie Blackburn, Xi Chen, and Yuan Zhang

System Overview



Improvements in Content Selection

Larger background corpus for LLR

- Half of the New York Times corpus on Patas

Tweaking MLP regression

- 1 hidden layer of size 50
- Adaptive learning rate

Sentence Compression

Little to no effect on scores (R2 -14%):

- Ages
- Dates/times
- Attributions

Negative effect on scores (R2 -26%):

- Adjectives
- Adverbs
- Initial Conjunctions

Modifications in Content Realization

Sentence compression is introduced

In content realization, a modified greedy algorithm is applied:

1, while **compressed sentence length** does not exceed word limit:

2, pick the sentence with the highest score among candidates

3, unless the sentence's tf-idf similarity with candidates exceed threshold

(t <0.4)

Info Ordering: Review

As the result, we augment each summary sentence into a sentence group in the input documents by label spreading. 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.

Experiments on Info Ordering

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

Name	tau	Description
Adjacency 1	0.39	dim(word vector) = 50
Adjacency 2	0.41	dim(word vector) = 100
Adjacency 3	0.44	dim(para2vec) = 100
Adjacency 4	0.33	dim(word vector) = 200
Chronological	0.46	

Reflection on Info Ordering

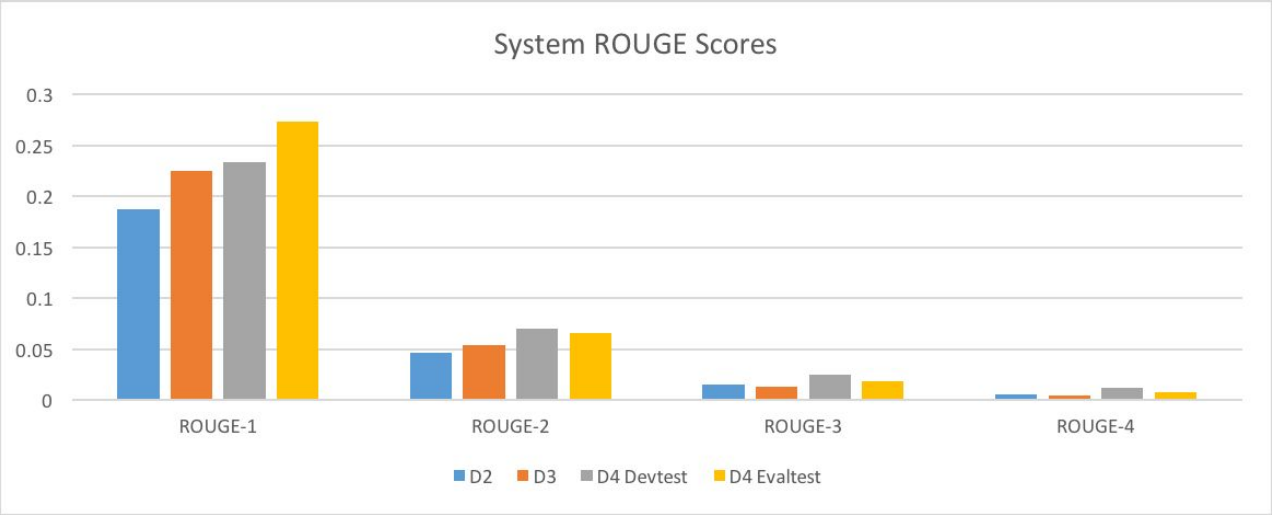
No word/sentence embedding based similarity can exceed chronological ordering. Still, they do better in different cases/test passages. Therefore, I assume a more efficient algorithm should take both into account. A more preferable solution is ordering/clustering by:

$$\text{Simlarity}(s_1, s_2) =$$

$$t_1 * \text{semantics\_diff}(s_1, s_2) + (1 - t_1) * \text{chonological\_diff}(s_1, s_2)$$

This is just a variant of Bollegala et al(2012), ‘A preference learning approach to sentence ordering for multi-document summarization’ - combine different criterion together with machine learning based parameter adjusting.

Results



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
D4 Devtest	0.23406	0.07048	0.02468	0.01198
D4 Evaltest	0.27357	0.06555	0.01845	0.00766

Issues and Successes

Minimal sentence compression has little to no effect on scores

Aggressive sentence compression has negative effects on scores and readability

Feature dependency

Ex. 1: “Parkinson’s disease”

D3

I saw it listed as the cause of death in an obituary. Exercise alone was enough to prevent the degeneration of brain cells in rats with **Parkinson**'s disease, University of Pittsburgh researchers report. Today, a major treatment strategy is aimed at developing medicines to stop this abnormal protein from clumping. The research also identified a group of brain receptors in mice that appear to be responsible for nicotine addiction. It is a form of dementia that the National Institute of Neurological Disorders and Stroke calls dementia with Lewy bodies. The loss of cells that produce the neurotransmitter dopamine causes the telltale tremors, rigid and slow movements of **Parkinson'**s. Distributed by the Los Angeles Times/Washington Post News Service

D4

Fox, who has **Parkinson**'s disease, campaigned with Kerry in New Hampshire on Monday and filmed the ad after the event. The 84-year-old has **Parkinson**'s disease, which makes it difficult for him to walk and to pronounce his words. Exercise alone was enough to prevent the degeneration of brain cells in rats with **Parkinson**'s disease, University of Pittsburgh researchers report. The loss of cells that produce the neurotransmitter dopamine causes the telltale tremors, rigid and slow movements of **Parkinson'**s. John Paul _ the most traveled pope in history _ cut back on his trips a few years ago. Investigators studied five families with a history of **Parkinson**'s disease who lived in the Basque region of Spain and in England.

Ex. 2: “bird flu”

D3

While Hong Kong's imports mostly come from Shenzhen, Macao's are mainly from Zhuhai, Zhongshan and Jiangmen cities. A SAR government spokesman said the group, which comprises representatives from various government departments, will hold its first meeting Tuesday. A 54-year-old succumbed to the virus early this month, but a 2-year-old boy recovered after hospitalization in November. Although several cases were reported in Hong Kong, none was found in Macao so far. Macao residents are eating less birds these days although there has been no report of bird flu cases in Macao. The sheet of paper provided information of Influenza A H5N1, cautioning tourists to keep up a good body immunity.

D4

The "bird flu" has claimed four victims here, killing two, including a 54-year-old man who died Friday. A SAR government spokesman said the group, which comprises representatives from various government departments, will hold its first meeting Tuesday. Seven people have been infected so far, with two dead and two in critical conditions. The poultry imported from China's inland areas are from a different source from those imported by Hong Kong, they said. Local chicken farmers say that sales have dropped between 30 percent to 50 percent over the past weeks. Although several cases were reported in Hong Kong, none was found in Macao so far.

Ex. 3: “mad cow disease”

D3

The human form of **mad cow** disease is called variant Creutzfeldt-Jakob. The fatal brain-wasting disease is believed to come from eating beef products from cows struck with **mad cow** disease. **Tons of meat went to the market in violation of European norms", Valchovski, a virus expert and medical doctor, told AFP.** The money would come from the euro188 million (US$235 million) set aside in 2005 to combat animal diseases in the EU. The vast bulk of them are elderly dairy cattle who would have eaten cattle-based feed in the 1980s. The likely vector of contamination for livestock was brain and nerve tissue mixed in animal feed.

D4

The human form of **mad cow** disease is called variant Creutzfeldt-Jakob. The fatal brain-wasting disease is believed to come from eating beef products from cows struck with **mad cow** disease. **Some 141 people are known to have died of vCJD in Britain.** Ireland banned the use of meat and bone meal as cattle feed, the suspected origin of **mad cow** disease, in 1990. It is estimated the **mad cow** crisis has cost the Canadian beef industry and rural economies about 5 billion US dollars. The money would come from the euro188 million (US$235 million) set aside in 2005 to combat animal diseases in the EU. It icope with this difficult situation," she said.

References

[1] Ji Donghong and Nie Yu. Sentence ordering based on cluster adjacency in multi-document summarization. *The third international joint conference on natural language processing*, 2008.

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

[3] 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.

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

[5] 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.

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