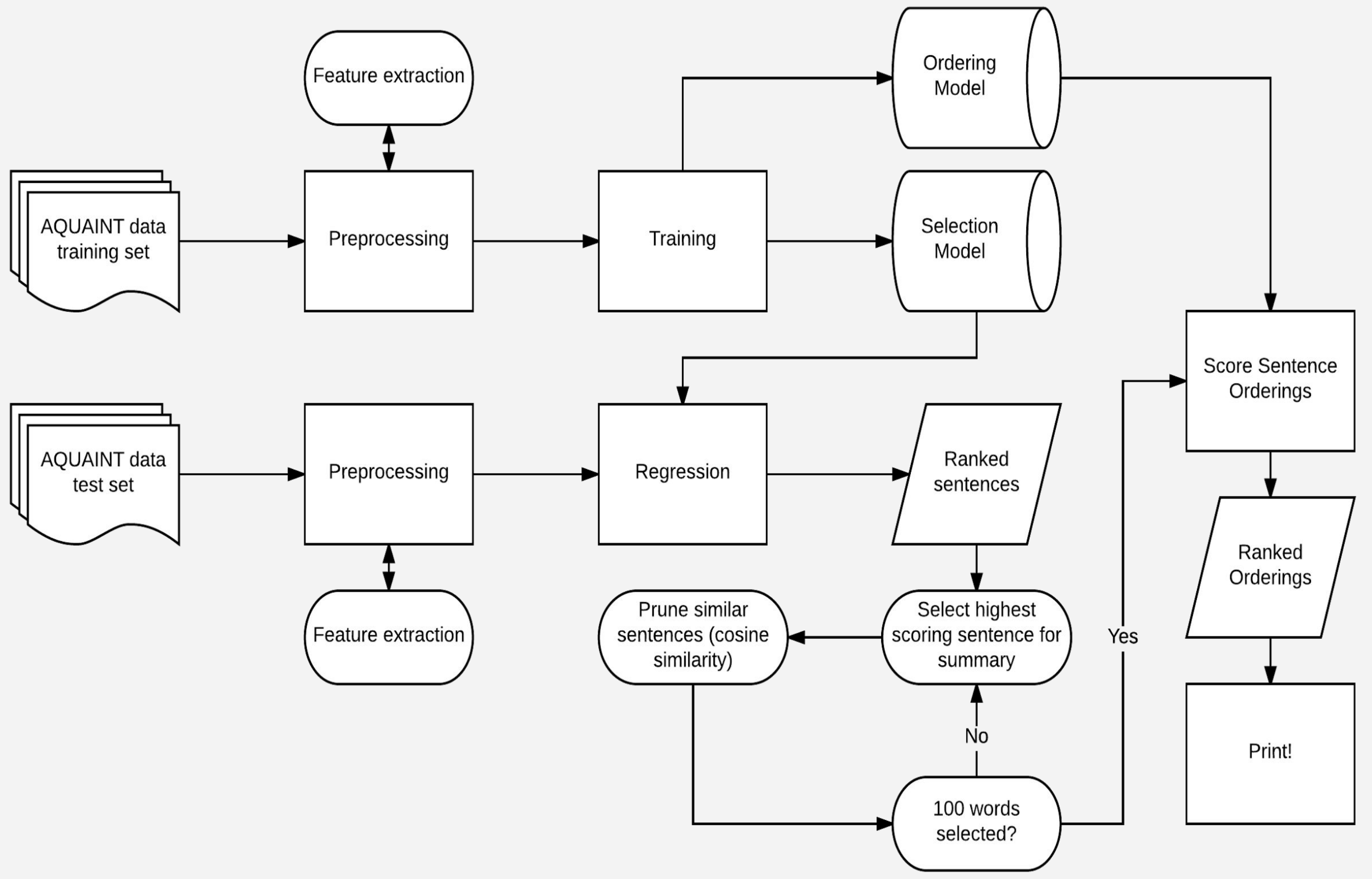
Deliverable #3

Alex Spivey, Eli Miller, Mike Haeger, and Melina Koukoutchos
May 18, 2017

System Architecture



Improvements in Content Selection

- Preprocessing
  - We removed boilerplate and other junk data
  - Split the sentences into two forms:
    - One that is lowercase and stemmed
    - And another that preserves its raw form for later use in building summaries
  - Added two new features
    - NER percentages
    - LexRank
- Gold Standard Data
  - Use cosine similarity to tag document sentences as in the summary

Improvements in Content Selection

- Features
  - Previously: TF-IDF, sentence position
  - New:
    - NER (named entities in sentence / length of sentence)
    - LexRank
    - Sentence length
- Similarity Measure
  - Cosine similarity (words stemmed and lowered)
    - Threshold testing

Information Ordering

- Based on a logistic regression model
  - Scores ordered pairs of adjacent sentences
  - Based on tf-idf scores of each sentence and similarity
- Overall score of an ordering:
  - Sum of scores of each pair
- Ordering with highest score selected

Sample Summaries

The four New York City police officers charged with murdering Amadou Diallo returned to work with pay Friday after attending a morning court session in the Bronx in which a Jan. 3 trial date was set. Marvyn M. Kornberg, the lawyer representing Officer Sean Carroll, said Thursday that in addition to standard motions like those for discovery _ in which lawyers ask prosecutors to hand over the information they have collected _ he expected defense lawyers to ask the judge to review the grand jury minutes to decide if the indictments were supported by the evidence.

"In terms of bio-diversity protection, Qinling and Sichuan pandas need equal protection, but it is a more urgent task to rescue and protect Qinling pandas due to their smaller number," Wang Wanyun, chief of the Wild Animals Protection section of the Shaanxi Provincial Forestry Bureau, told Xinhua. On Dec. 14 last year, Feng Shiliang, a farmer from Youfangzui Village, told the Fengxian County Wildlife Management Station that he had spotted an animal that looked very much like a giant panda and had seen giant panda dung while collecting bamboo leaves on a local mountain.

Results

ROUGE Recall

	D2	D3
ROUGE-1	**0.18765**	0.16459
ROUGE-2	**0.0434**	0.03768
ROUGE-3	0.01280	**0.01289**
ROUGE-4	0.00416	**0.00439**

- Best combination of features: Sentence length and position
- TF-IDF and/or LexRank?

Issues & Successes

- Issues:
  - What is an ideal number of gold standard sentences to tag?
  - Why aren’t certain features improving content selection?
  - ROUGE-1 and ROUGE-2 decreased
- Successes:
  - Gold standard data problem from D2 addressed
  - Information ordering implemented
  - ROUGE-3 and ROUGE-4 improved slightly

Future Improvements

- TF-IDF similarity
- More threshold testing (gold standard data, content selection)
- New features for information ordering
- Feature combination testing (content selection, information ordering)
- Prune negative examples to get more balanced positive/negative training split
- Content realization

Resources

Meng Wang, Xiaorong Wang, Chungui Li and Zengfang Zhang. 2008. *Multi-document Summarization Based on Word Feature Mining*. 2008 International Conference on Computer Science and Software Engineering, 1: 743-746.

You Ouyang, Wenjie Lia, Sujian Lib, and Qin Lu. 2011. *Applying regression models to query-focused multi-document summarization*. Information Processing Management, 47(2): 227-237.

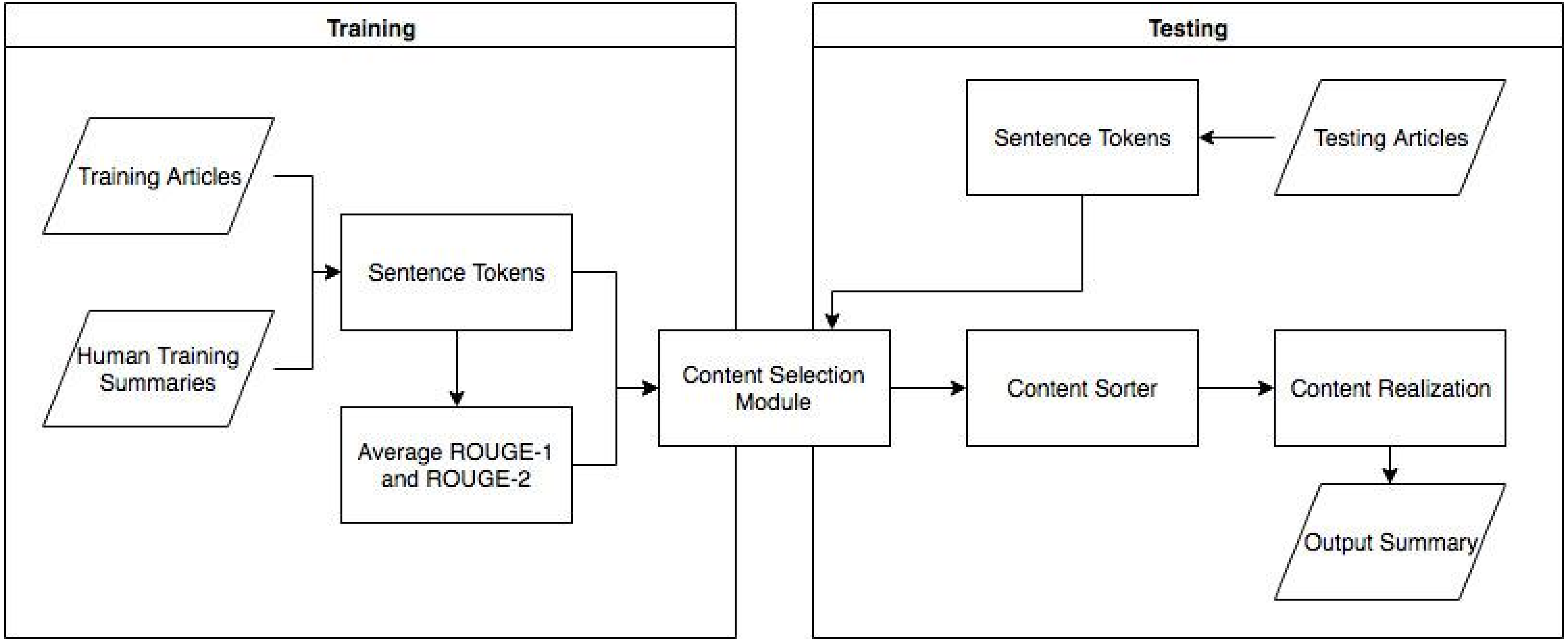
Günes Erkan and Dragomir Radev. 2004. *Lexrank: Graph-based lexical centrality as salience in text summarization*. Journal of Artificial Intelligence Research, 22:457–479.

Sandeep Sripada, Venu Gopal Kasturi, and Gautam Kumar Parai. 2005. *Multi-document extraction based Summarization*. CS224N Final Project. Stanford University.

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)

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

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:

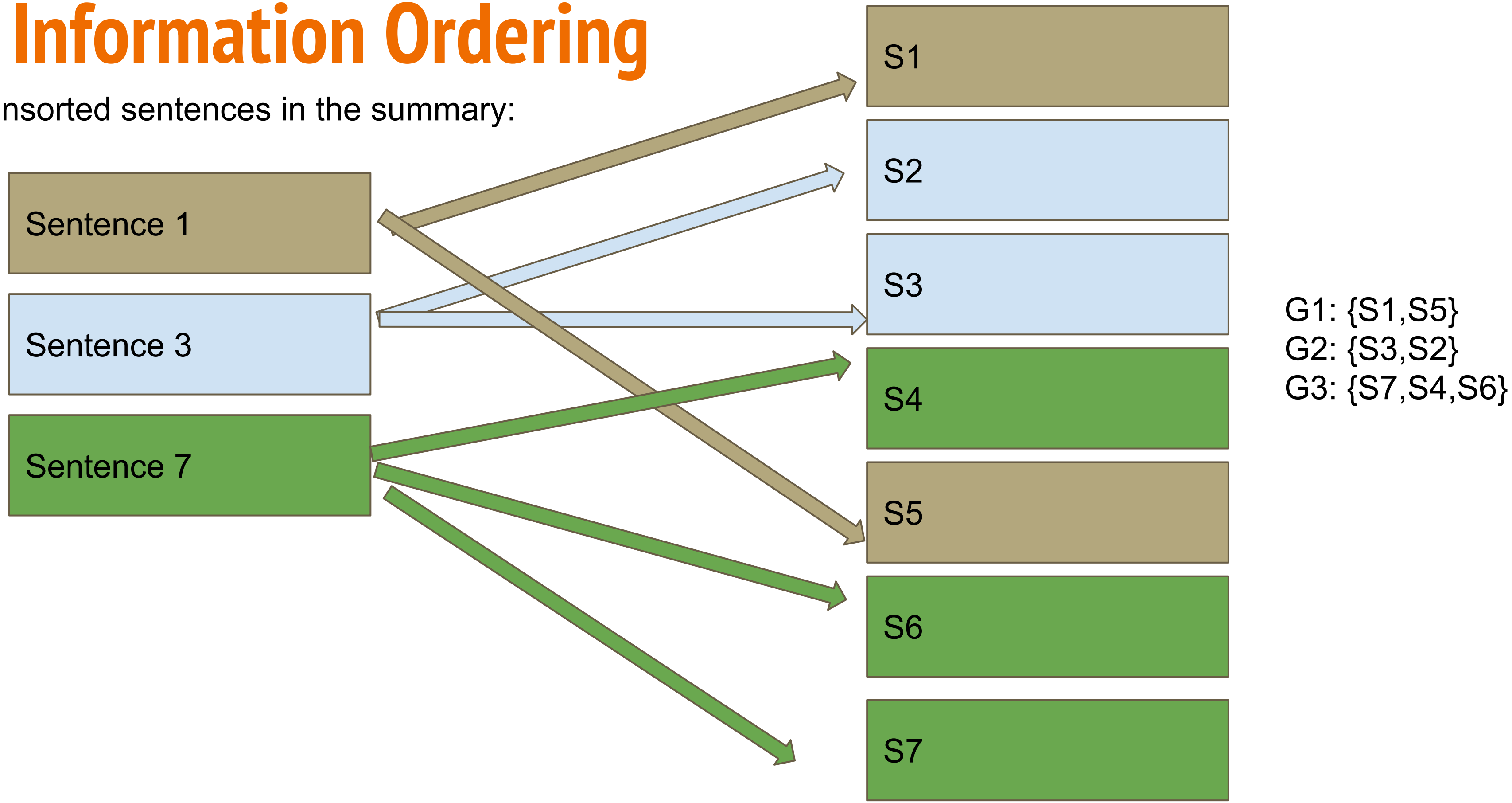
$$C_{m,n} = f(G_m, G_n)^2 / (f(G_m)f(G_n))$$

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

ordered sentences in original documents:

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 (symmetric window size = 2):	0.200
Adjacency (symmetric window size = 1):	0.324
Adjacency (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.

Automatic Summarization System

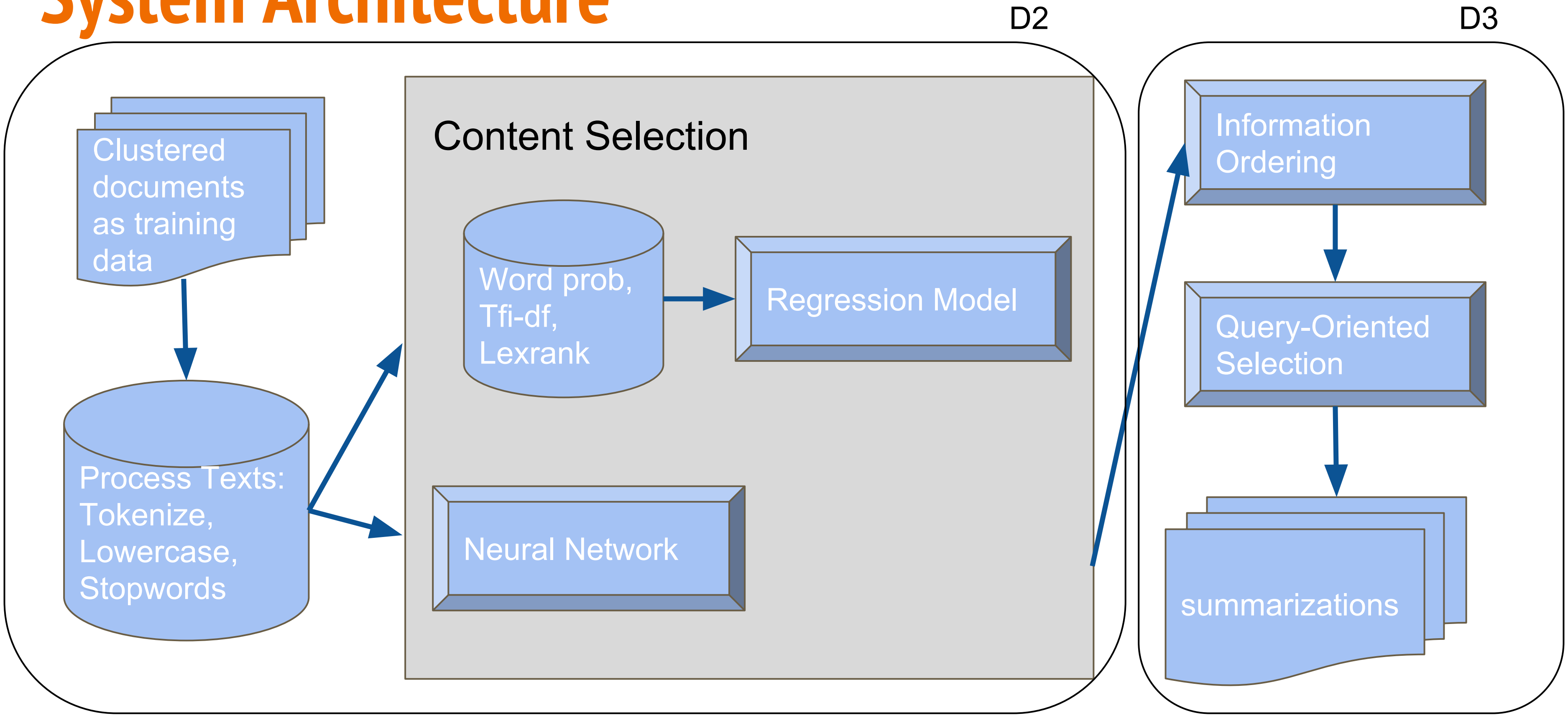
DELIVERABLE 3: Information Ordering & Topic-focused Summarization

Wenxi Lu, Yi Zhu, Meijing Tian

Outline

- System Architecture
- Baseline
- Information Ordering
- Topic-focused Summarization
- Results
- Issues and Discussion

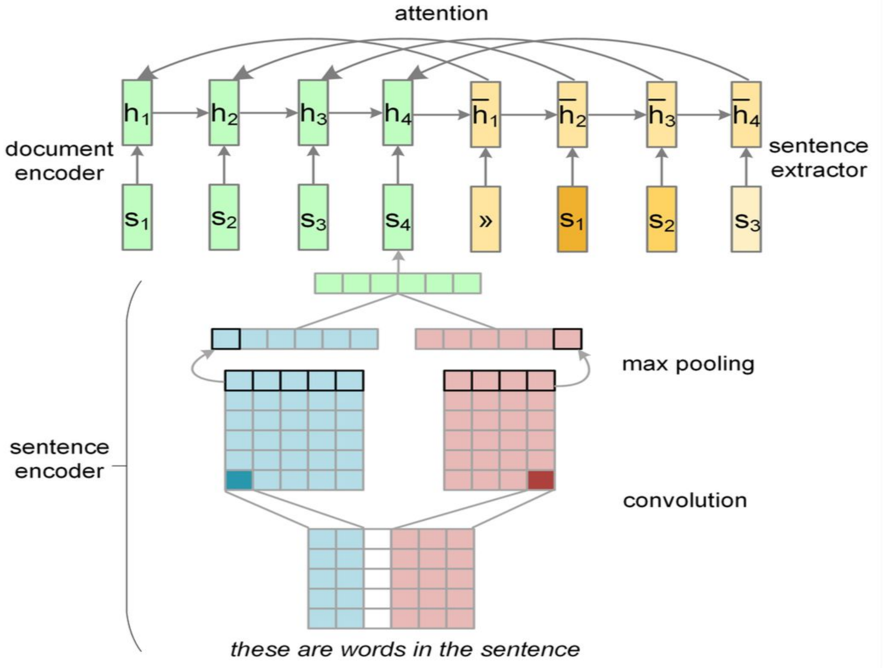
System Architecture



Baseline

$$\bar{\mathbf{h}}_t = \mathrm{LSTM}(p_{t-1}\mathbf{s}_{t-1}, \bar{\mathbf{h}}_{t-1})$$

$$p(y_L(t) = 1|D) = \sigma(\mathrm{MLP}(\bar{\mathbf{h}}_t : \mathbf{h}_t))$$



- Changes
  - Training with scheduled sampling

    $$p_{t-1} = \alpha p^d_{t-1} + (1 - \alpha) y_{t-1}$$
  - Output first n sentences with label 1
    - Criterion
      - n not too small
      - Higher Precision
    - Output all sentences with label 1
  - Format
    - New line split doc summaries
    - Summaries sorted by date

Neural Summarization by Extracting Sentences and Words [Cheng et al; 2016]

Information Ordering

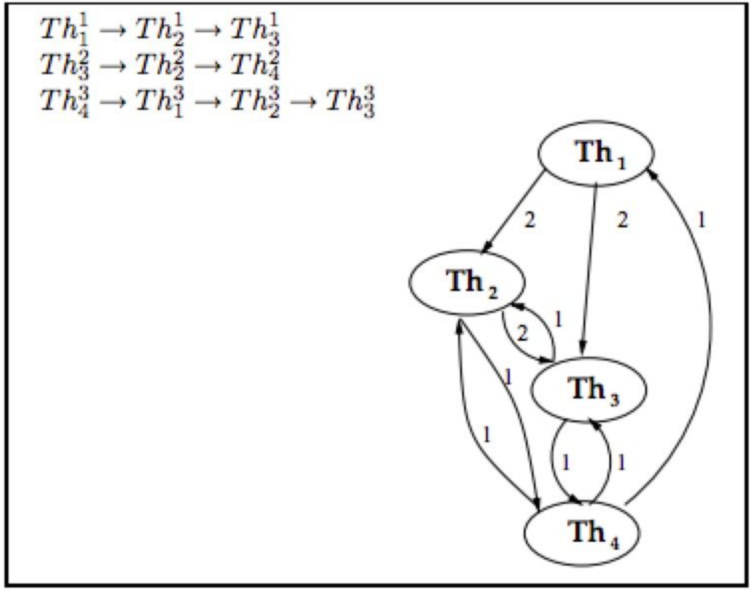
- Sentence Clustering
- Majority Ordering
- Chronological Ordering

Sentence Clustering

- Sentence Clustering API by **rxNLP**
- http://www.rxnlp.com/api-reference/cluster-sentences-api-reference/
- **Input**: text of sentences or list of sentences
- **Output**: dictionaries, with each value being another dictionary represents the information of a cluster that are similar to each other.

Majority Ordering

- A linear ordering between themes which maximizes the agreement between the orderings provided by the input texts
- Algorithm: Barzilay et al, 2002
- A modified version of topological sort



Downside?

Th_i^j is the sentence part of the theme i in the input ordering j.

Weights = the sum of the weights of its outgoing edges minus the sum of the weights of its incoming edges

Initial weights:

Weight_1 = 2 + 2 - 1 = 3

Weight_2 = 2 + 1 - 1 - 1 = 1

Weight_3 = 1 + 1 - 2 - 1 = -1

Weight_4 = 1 + 1 + 1 - 1 - 1 = 0

Chronological Ordering

- Algorithm: Barzilay et al, 2002
- Articles published on different dates
- Approximate the theme time by its first publication time
- Input: a set of themes and orders them chronologically
- If there is a tie in publishing data: sort them according to their order of presentation in this article

Topic-focused Summarization

- Query-oriented lexrank score
- Extends lexRank, finding sentences that are more similar to the query while carrying the most information.
- Algorithm: Otterbacher 2005

Query-oriented LexRank

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

rel(s|q) is the relevance of a sentence given a query,
d referred as “question bias,” is a trade-off between two terms

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

Results

	R1	R2	R3	R4
NN (D2)	0.22868	0.05655	0.01540	0.00394
Baseline	0.2079	0.0603	0.02079	0.00837
Baseline + CO	0.21740	0.05778	0.01813	0.00597
Baseline + MO	0.15813	0.03193	0.00837	0.00244
Baseline + CO + lexRank	0.18886	0.04335	0.01297	0.00480
Baseline +MO + lexRank	0.1743	0.0387	0.01178	0.0041

Examples

1st lead: islamic group says killed hariri due to ties with saudi arabia: al-jazeera
a previously unknown islamic group on monday claimed responsibility for an earlier killing of former lebanese prime minister rafik hariri due to his ties with saudi arabia , the qatar-based al-jazeera tv channel reported .
opposition blames syrian and lebanese governments for hariri's assassination
opposition leaders on monday held lebanese and syrian governments responsible for the assassination of former prime minister rafik hariri , demanded syrian troops withdraw from lebanon within the next three months and called on the international community to intervene to help `` this captive nation . ''
`` we hold the lebanese authority and the syrian authority , being the authority of tutelage in lebanon , responsible for this crime and other similar crimes , '' said a statement after an opposition meeting held monday night at the late leader 's house in beirut .
…...

1st lead: islamic group says killed hariri due to ties with saudi opposition leaders on monday held lebanese and syrian governments responsible for the assassination of former prime minister rafik hariri , demanded syrian troops withdraw from lebanon within the next three months and called on the international community to intervene to help `` this captive nation . ''

Baseline R1: 0.220, R2: 0.099

Baseline + Mo + lexRank R1: 0.342, R2: 0.121

Issues & Discussion

- Checking the implementation
- Improving the sentence cluster
- Tuning the parameters
- Trying different approaches
- Improving the outputs produced by NN

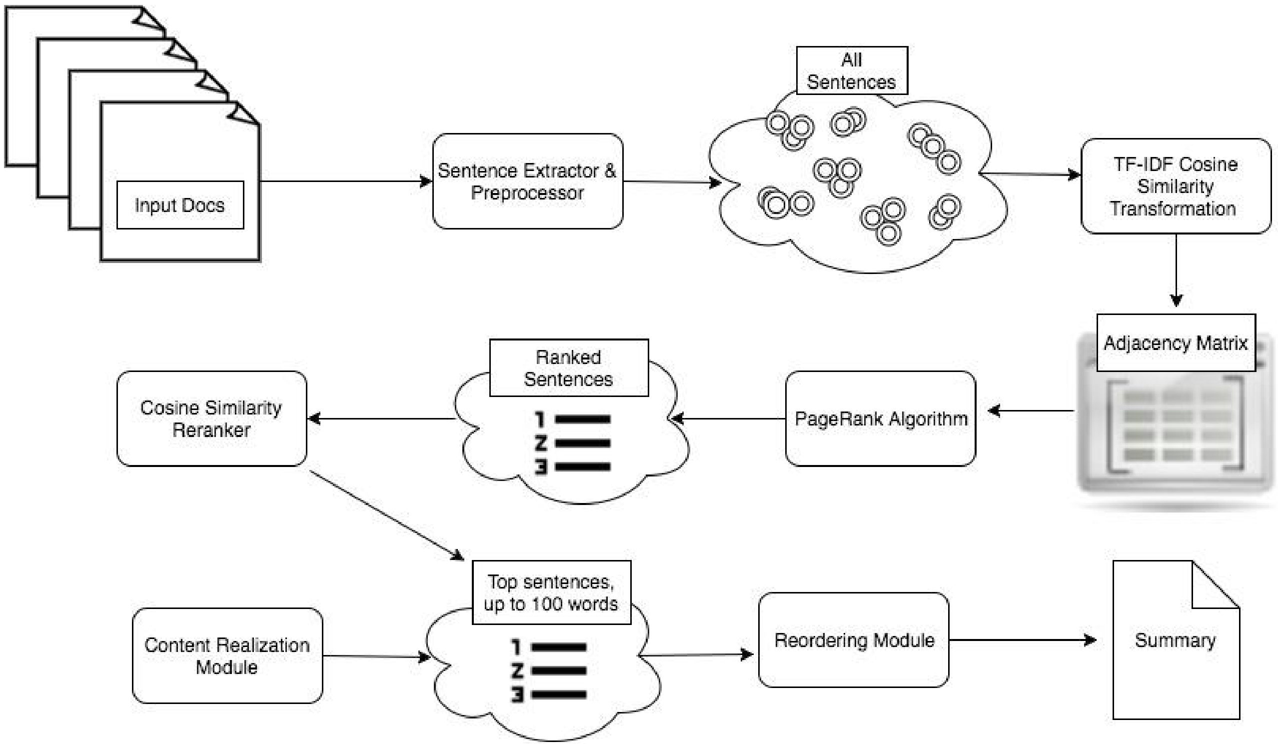
References

- Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multi-document news summarization. Journal of Artificial Intelligence Research 17:35–55.
- Gunes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22:457–479.
- Jahna Otterbacher, Gunes Erkan, and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. Journal of Artificial Intelligence Research .
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28*, pages 1171–1179. Curran Associates, Inc.

Multi-Document Summarization - D3

Eslam Elsawy, Audrey Holmes, Masha Ivenskaya

System Architecture



Improvements to Content Selection

Filtering Criteria that Improved System Performance:

- Remove sentences with phone numbers or websites
- Remove location phrases, e.g. CHICAGO, Illinois (AP) --
- Remove sentences shorter than 35 characters

Filtering Criteria that Did NOT Improve System Performance:

- Remove abbreviations, e.g. Papua New Guinea **(PNG)**
- Remove ages, e.g. Ana, **36**

Content Selection: Lemmatization and Stemming

WordNet Lemmatizer

- Did not improve system performance, seemed to have insufficient coverage

Porter Stemmer

- Did not improve system performance, seemed to over-stem words

Snowball Stemmer

- Improved system performance, good coverage and less over-stemming

Content Selection: Binary tf-idf

t = term, d = document, D = corpus

Original term frequency: $f_{t,d}$ = # of times term t appears in document d

Binary term frequency: $f_{t,d}$ = {1 if term t appears in document d; 0 otherwise}

$$\frac{f_{t,d}}{\log \frac{|D|}{1 + |d \in D : t \in d|}}$$

Content Selection: Doc2Vec/GloVe

We continued to experiment with using more sophisticated methods of computing sentence similarity for LexRank:

- Trained a **doc2vec** model on the full Reuters corpus: Rouge scores lower than with tf-idf cosine similarity.
- Used pre-trained **GloVe** embeddings and manually calculated sentence vectors: Rouge scores lower than with tf-idf cosine similarity.

Reordering Approaches - Chronological

Sentences ordered by:

- Publication date/time
- If same, by order in original document

PRO: Worked well for topics about events unfolding over time, like natural disasters.

CON: In many cases summaries lack cohesion

Chronological Reordering Examples

“the papua new guinea (png) defense force, the police and health services are on standby to help the victims of a tsunami that wiped out several villages, killing scores of people, on png's remote north-west coast friday night. igara said reports so far indicated that a community school, government station, catholic mission station and the nimas village in the sissano area west of aitape had been completely destroyed, where 30 people were dead.   the death toll in papua new guinea's (png) tsunami disaster has climbed to 599 and is expected to rise, a png disaster control officer said sunday."

“for example, new roads will be banned in national forests around the park, servheen said. fish and wildlife service is poised to remove the park's renowned bears from the endangered species list. federal wildlife officials estimate that more than 600 grizzly bears live in the region surrounding yellowstone in idaho, montana and wyoming. grizzly bears in and around yellowstone national park should be removed from the endangered species list after 30 years of federal protection, the u.s. department of interior said tuesday. the only other large population of grizzlies in the united states is in and around glacier national park."

Reordering Approaches - (naive) Cohesion

- All permutations of sentences are created and assigned a cohesion score
- Cohesion score: sum of cosine similarity scores of adjacent sentences
- The order with the highest cohesion score is chosen

PRO: Sentences “link” together well in most summaries

CON: First sentences often bad, chronology often skewed.

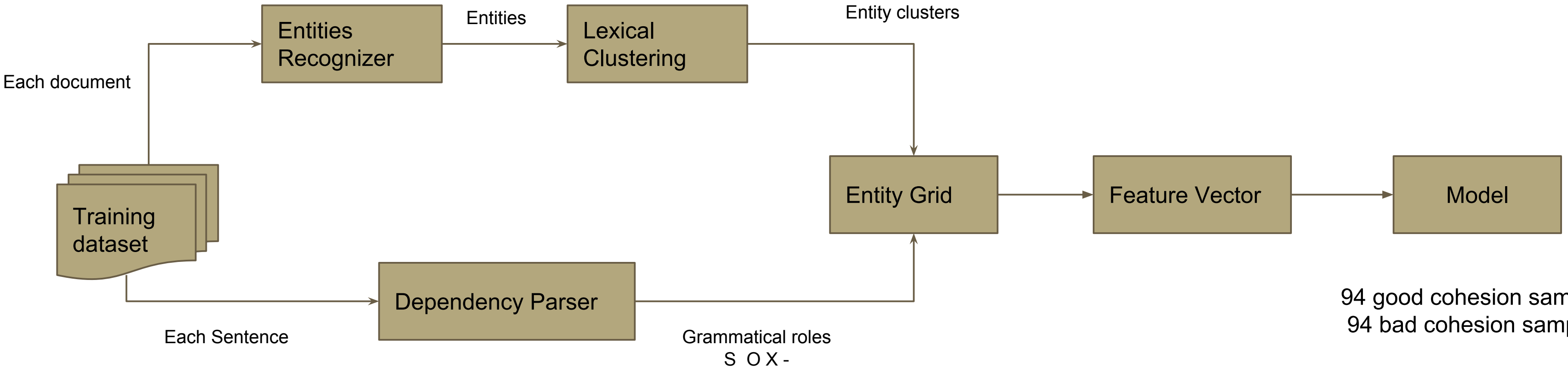
(Naive) Cohesion Reordering Examples

“in the united states, 21 percent of known species are threatened or extinct. the survey, published online by the journal science, studied the 5,743 known amphibian species and found that at least 1,856 of them face extinction, more than 100 species may already be extinct, and 43 percent are in a population decline many for unknown reasons. the researchers called for efforts to protect the habitat of amphibians and to reproduce the threatened species in captivity. habitat decline, from deforestation to water pollution and wetlands destruction, threatens them because the animals live both on land and in water."

“burke was in the family's boulder home when 6-year-old jonbenet was found beaten and strangled dec. 26, 1996. hunter took the jonbenet case to the grand jury shortly after a former boulder police detective on the case and three former friends of the ramseys publicly demanded that colorado's governor, roy romer, replace hunter on the case with a special prosecutor. although the police chief and district attorney both have said that the ramseys fall under ``the umbrella of suspicion,'' they have not formally named any suspects. police say her parents, john and patsy ramsey, remain under suspicion. "

Reordering Approaches - Entity Grid Cohesion

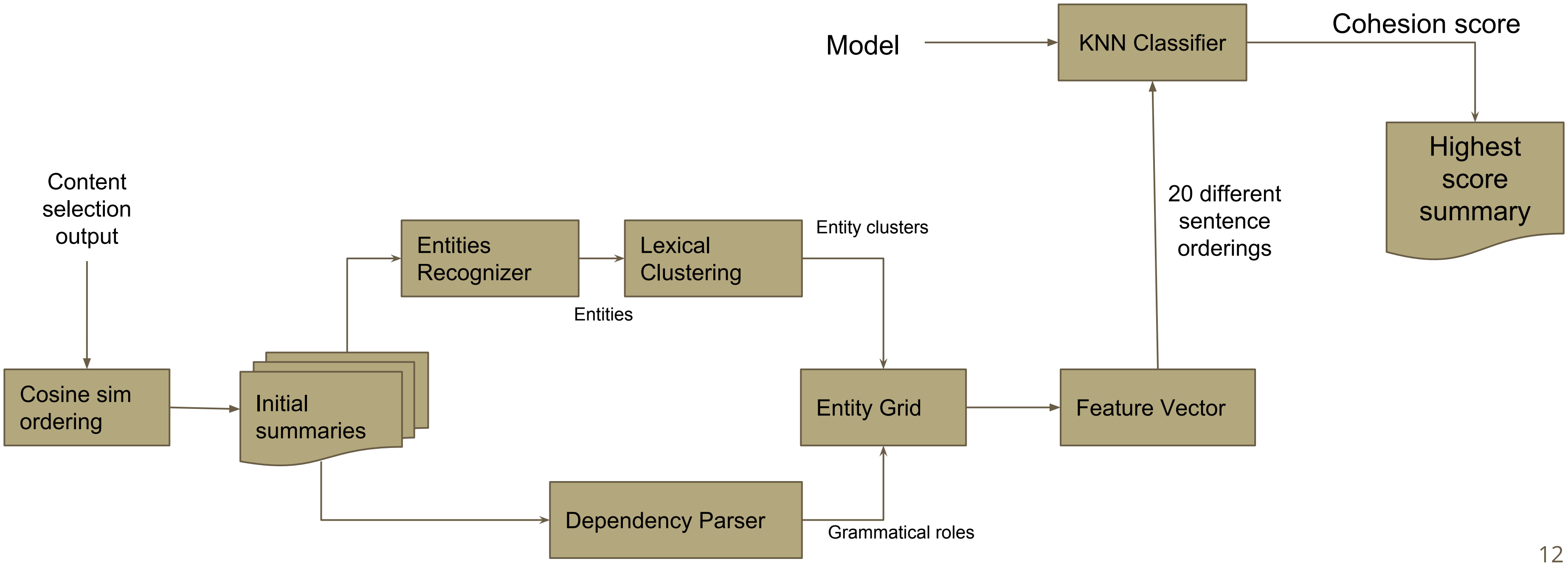
Offline training:



94 good cohesion samples
94 bad cohesion samples

Reordering Approaches - Entity Grid Cohesion - II

Run time:



Entity Grid Ordering Examples - I

Vioxx Drug Announcement

Id	Sentence
1	the announcement comes just two weeks after merck pulled its painkiller, vioxx, which is in the same class of drugs as bextra, from the market because a study showed that the risk of heart attacks doubled for patients who had taken vioxx for 18 months or longer.
2	vioxx was approved after trials held under the auspices of the food and drug administration showed it to be effective (which it was).
3	gilmartin was clear that the trial should be halted and that the drug might have to be taken off the market.
4	the drug was not pulled at that point.

Initial Ordering: 1, 2, 3, 4
Score: 0.55

Best Ordering: 2, 1, 4, 3
Score: 0.73

Entity Grid Ordering Examples - II

Columbine school shooting

Id	Sentence
1	columbine is about three miles away
2	tuesday morning 12 columbine high school students and a teacher were murdered when eric harris and dylan klebold, also columbine students, opened fire with at least four guns and dozens of bombs
3	students returned to classes thursday at chatfield high school, but the bloodbath at rival columbine high haunted the halls
4	the nation and the world have joined in grieving for the students of columbine, gore said
5	harris and klebold, who authorities say planned the massacre for more than a year, have been portrayed by classmates as outcasts from the popular students at columbine

Initial Ordering: 1, 2, 3, 4, 5
Score: 0.45

Best Ordering: 2, 3, 1, 5, 4
Score: 0.64

Entity Grid Ordering - Issues and Successes

Success:

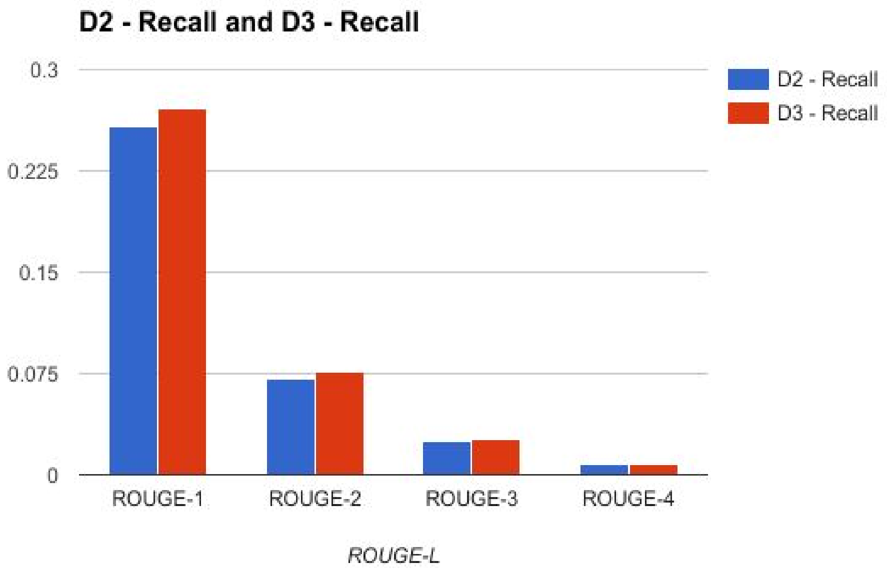
- Offline Training and Online Testing procedures
- Promising output

Issues:

- Lack of evaluation metric
- A lot of tuning parameters:
  - Different training set sizes
  - Different classifiers
  - Number of different orderings considered

Rouge Scores

ROUGE-L	D2 - Recall	D3 - Recall
ROUGE-1	0.25785	0.27056
ROUGE-2	0.07108	0.07684
ROUGE-3	0.02438	0.02596
ROUGE-4	0.00847	0.00739



Next Steps

Information Ordering:

- Try mix different ordering techniques like experts then entity grid
- Have an evaluation metric and tune parameters

Content Realization:

- Fix grammatical errors, (e.g. sentence fragments, misplaced clauses, etc.)
- Replace redundant proper nouns with appropriate pronouns
- Sentence Compression

References

[1] Radev, Dragomir R., et al. "MEAD-A Platform for Multidocument Multilingual Text Summarization." *LREC*. 2004.

[2] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research 22 (2004): 457-479.

[3] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. 2004.

[4] Barzilay, Regina, and Mirella Lapata. "Modeling local coherence: An entity-based approach." *Computational Linguistics* 34.1 (2008): 1-34.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370

[6] Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Questions ?