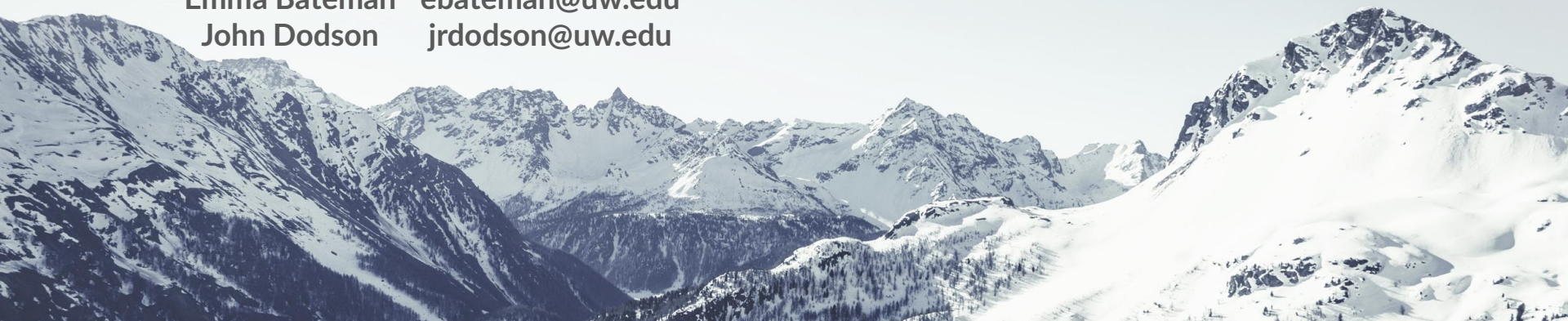
LING 573 Project

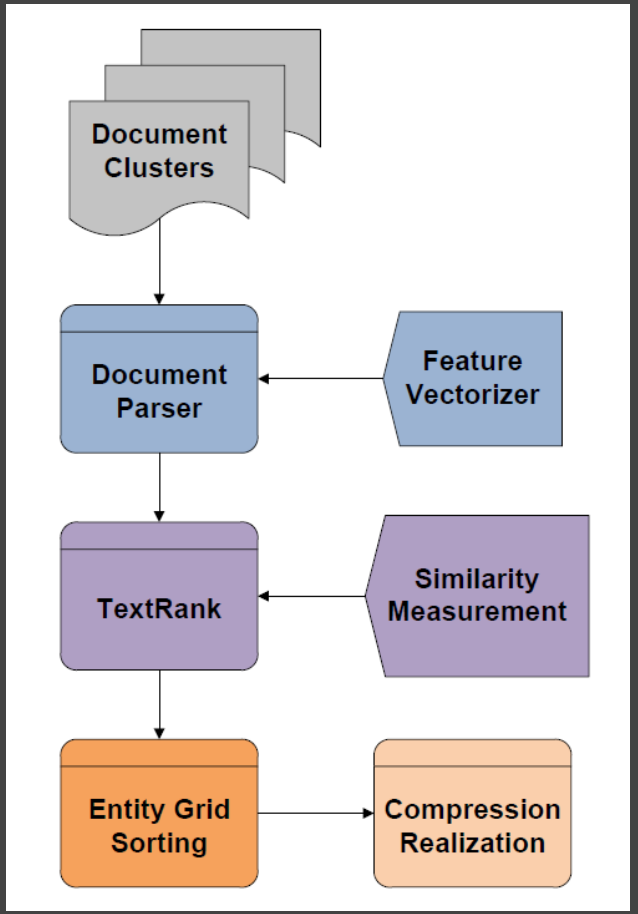
Charlie Guo qcg@uw.edu
Emma Bateman ebateman@uw.edu
John Dodson jrdodson@uw.edu



Overview

- Baseline extractive system
- Graph-banked sentence ranking
- Entity grid information ordering
- Sentence compression

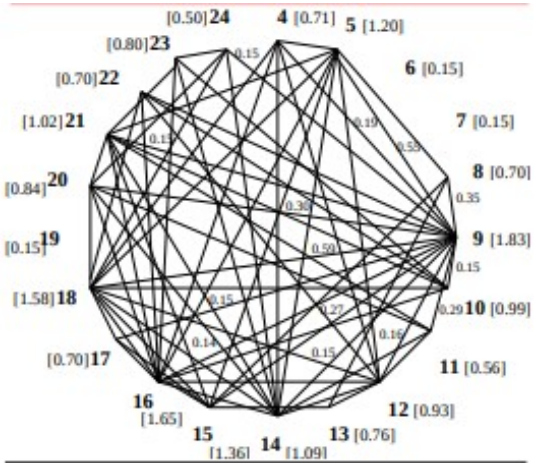
System Architecture




Approach

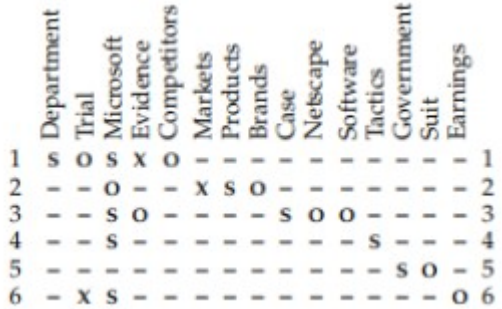
Content Selection

- Sentence saliency estimated with TextRank
- Sentences converted to unigram feature vectors
- Ranked from highest to lowest relevance
- Duplicates are removed
- Sentences selected until word count is met



Information Ordering: Entity Grid

- Based on Barzilay and Lapata (2008)
- Find entities mentioned in summary
- For each entity, sentence pair determine grammatical role if any
- Assign weights: 1.0 for subject, 0.5 for object, 0.1 for none
- Calculate weight for each entity over all candidate sentences
- Pick highest weighted entity and select sentence with the highest weight for that entity
- Repeat until all sentences are sorted



	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	s	o	s	x	o	–	–	–	–	–	–	–	–	–	–	1
2	–	–	o	–	–	x	s	o	–	–	–	–	–	–	–	2
3	–	–	s	o	–	–	–	–	s	o	o	–	–	–	–	3
4	–	–	s	–	–	–	–	–	–	–	–	s	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	s	o	–	5
6	–	x	s	–	–	–	–	–	–	–	–	–	–	–	o	6

Information Ordering: Shortest Path

- Generate bag-of-words feature vector for each sentence
- Find cosine distances
- Brute force search to generate shortest path
- More salient endpoint chosen as starting point

Content Realization

- Sentence compression
- Named Entity simplification

Sentence Compression: Rule-based

Tree-based PP Removal

- Obtain a syntax tree of a sentence
- Identify PP and check removability
- Realize sentence without PP

Example:
Senior Palestinian official Yasser Abed Rabbo denied on Tuesday reports saying that Palestinian leader Yasser Arafat has died ~~in a French hospital~~.

Sentence Compression: Rule-based

Rule-based Datetime removal

- Identify datetime entities
- Flag entities as “Removed”
- Identify preceding prepositions and flag
- Realize

Example:

Senior Palestinian official Yasser Abed Rabbo denied ~~on Tuesday~~ reports saying that Palestinian leader Yasser Arafat has died in a French hospital.

Sentence Compression: Trained Model

Training Data

- BBC newswire sentences with hand-written compressions
- Originally used in Clarke and Lapata (2008)
- Downloaded from: jamesclarke.net/research/resources

MaxEnt model

- 1000 training sentences parsed
- Each node of tree labeled “keep” (1) or “delete” (0)
- Training features such as: POS, sister node POS, depth, conatins stopwords, contains punctuation,

Sentence Compression: Trained Model

Training Data

- BBC newswire sentences with hand-written compressions
- Originally used in Clarke and Lapata (2008)
- Downloaded from: jamesclarke.net/research/resources

MaxEnt model

- 1000 training sentences parsed
- Each node of tree labeled “keep” (1) or “delete” (0)
- Binary classifier trained with sklearn Logistic Regression model

Sentence Compression: Trained Model

Training Features

- POS tag
- Sister node POS tag
- Within first three words?
- Within last three words?
- Contains word with >10 letters?
- Contains uppercase?
- Contains negation?
- Contain stopwords?
- Parenthetical?
- Lead adverbial?
- Lead preposition?
- Relative clause?
- Is root node?

Sentence Compression: Trained Model

Beam Search

- Keep/delete labels chosen using MaxEnt classifier and beam search
- Recursive search
- For each node, run beam search on each child node. Calculate probabilities for each resulting combination, plus probability of deleting current subtree. Return N most likely candidates.
- At leaf nodes, return [[0], [1]]

Sentence Compression Tools

- NLTK
- Stanford parser
- Scikit learn
- Clarke and Lapata (2008) compression data

Name Simplification

Examples

- Senior Palestinian official Yasser Abed Rabbo denied on Tuesday reports saying that Palestinian leader Yasser Arafat has died in a French hospital.
- As Arafat had struggled for life, there has been wild guess as to where he might be buried and where to hold the funeral service.
- Sunday night, the French foreign minister, Michel Barnier, told LCI television that Arafat was alive but that his circumstances were complicated.
- Palestinian leader ~~Yasser~~ Arafat would be buried at his headquarters in the West Bank town of Ramallah, well-informed Palestinian sources said Tuesday.

Name Simplification

Algorithm (on ordered sentences)

- Create a dictionary mapping full name to last name
- For each sentence, identify PERSON entities
- If a PERSON entity is a full name, replace it with last name
- Otherwise add this full name to the dictionary.

Name Simplification

Tools

- Spacy
- Python 3.6+

Results: Dev test

	ROUGE-1 recall	ROUGE-1 precision	ROUGE-1 F1	ROUGE-2 recall	ROUGE-2 precision	ROUGE-2 F1
Entity grid, rule-based	0.21643	0.23497	0.22322	0.05036	0.05311	0.05133
Shortest path, rule-based	0.21865	0.23280	0.22400	**0.05054**	0.05282	0.05136
Entity grid, trained model	0.20716	0.27003	0.23264	0.04610	0.05970	**0.05169**
Shortest path, trained model	**0.20730**	**0.27023**	**0.23280**	0.04586	**0.06043**	0.05173

Results: Eval test

	ROUGE-1 recall	ROUGE-1 precision	ROUGE-1 F1	ROUGE-2 recall	ROUGE-2 precision	ROUGE-2 F1
Entity grid, rule-based	0.25371	0.27168	0.26009	0.06173	0.06545	0.06281
Shortest path, rule-based	0.25628	0.26949	0.26154	0.06235	0.06441	0.06305
Entity grid, trained model	**0.25656**	**0.32682**	**0.28385**	**0.06695**	**0.08390**	**0.07365**
Shortest path, trained model	0.24993	0.31408	0.27512	0.06276	0.07798	0.06882

Related Readings

Rada Mihalcea and Paul Tarau, TextRank: Bringing Order into Texts, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004.

Regina Barzilay and Mirella Lapata, Modeling Local Coherence: An Entity-Based Approach Computational Linguistics, 2008

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations

James Clarke and Mirella Lapata, 2008. Global Inferece for Sentence Compression: An Integer Linear Programming Approach. In *Journal of Artificial Intelligence Research*, vol 31, pages 399-429.

Lu Wang et al. A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1384-1394. 2013

Thank you.