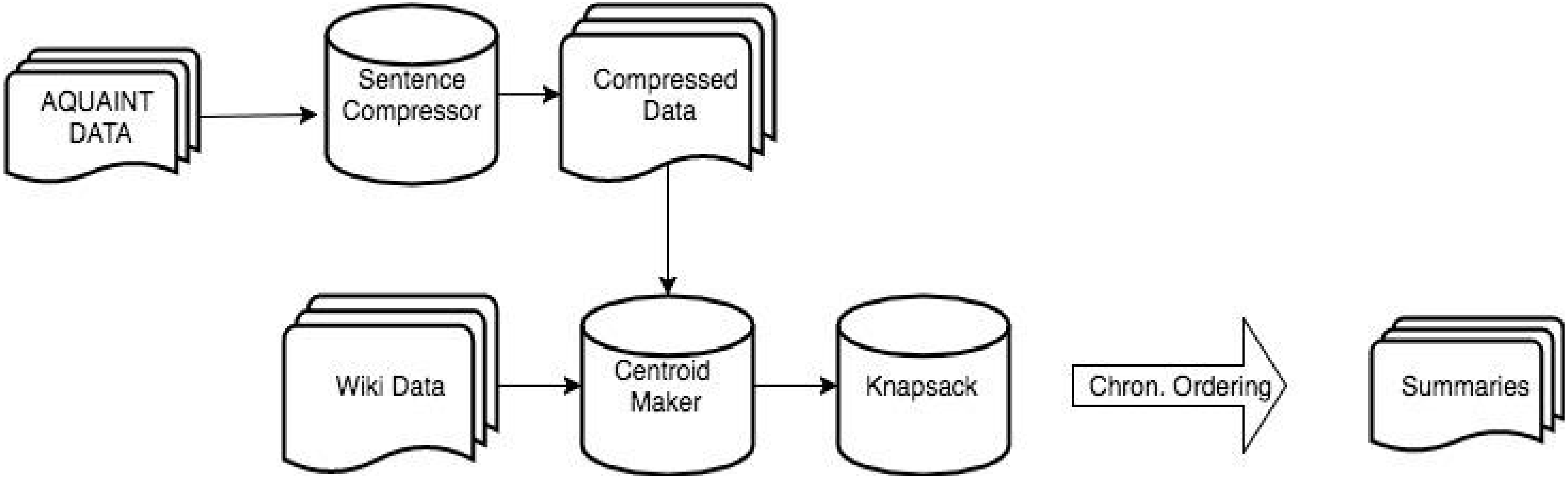
West Coast Python

Karen Kincy, Tracy Rohlin, Travis Nguyen

System Architecture:



Improvements:

- Sentence compression (Tracy)
- Improved sentence ordering (Tracy)
- Improved sentence cleaning (Karen and Tracy)
- Cosine similarity redundancy reduction (Karen)
- Wikipedia topic focus (Karen)
- Wikipedia background corpus for IDF and CBOW model (Karen)
- Optimization (Karen and Tracy)

Sentence Compression

Followed Zajic’s algorithm for sentence compression:

(1) Remove temporal expressions

- Removed things like days of the week, months, plus checked for “last”, “next”, “past”, “this” etc. within a 1+ word window
- Removed all adverbs except directionals (up/down, east/west…) as well as “virtually”, “allegedly”, “nearly”, “almost”

~~(2) Select Root S node~~

Sentence Compression...

(3) Remove preposed adjuncts

- Simple regex looking for a 2-3 words followed by a comma:
  - In summary, in conclusion, etch.
- Remove attributives
  - "..., the state reported.", "..., the judge ruled.", "..., he said."

~~(4) Remove some determiners~~ (reduces readability/grammaticality)

(5) Remove conjunctions;

- Keep conjuncts of 'but' but remove second conjunct of 'and'

~~(6) Remove modal verbs~~ (removed 'have' and 'can', but not others due to grammaticality)

Sentence compression…

~~(7) Remove complementizer that~~ (reduces readability/grammaticality)

~~(8) Apply the XP over XP rule~~ (XP doesn't seem to be part of the penn treebank node list)

Sentence Compression...

(9-15) Remove various SBARs and PPs

The chesapeake bay foundation led a rally in which speakers accused government officials of dragging their feet on bay cleanup measures.

The Chesapeake Bay foundation led a rally.

But...

Colonel James Pohl, halted proceedings after england indicated that she did not believe her actions were wrong.

Colonel James Pohl, halted proceedings after England indicated.

Redundancy vs. Relevance

- Cosine similarity redundancy reduction
  - More aggressive pruning after choosing topN sentences
  - Compare each vectorized sentence with every other sentence
  - Threshold of 0.7 optimizes ROUGE scores
- Wikipedia background corpus
  - Allows a finer representation of relevancy
  - IDF for TF*IDF calculation
  - CBOW model for Wikipedia topic focus score
- Wikipedia query expansion
  - Improved topic focus

Wikipedia Background Corpus

- Used tagged and cleaned Wikipedia corpus on Patas
  - About 67 GB total

  -
    ```
    egrep "#s-doc|#s-sent" /corpora/tc-wikipedia/wikipedia-tagged2_1.txt
    > wikipedia_sents.txt
    ```
  - Reduces size to under 11 GB
- Parsed first 50,000 articles
  - Saved IDF scores for terms
  - Trained CBOW model and cached out

CBOW Model

- CBOW (Continuous Bag of Words) model from Word2Vec
  - Trained on 50,000 documents from Wikipedia corpus
- Best training parameters:
  - cbow = Word2Vec(sentences, size=100, window=5, min_count=2, max_vocab_size=25000)
- CBOW model used to calculate similarity between terms
  - Building upon Tracy's topic focus score from D3
  - Used for Wikipedia topic focus score in D4
  - Compare similarity between embeddings rather than exact strings

Wikipedia Topic Focus

- Old approach: topic strings in devtest and evaltest
  - For example: “Cyclone Sidr”
  - Use CBOW model embeddings for “Cyclone” and “Sidr”
  - Check similarity with terms from candidate sentence
- Why not look up “Cyclone Sidr” in Wikipedia?
  - https://en.wikipedia.org/wiki/Cyclone_Sidr
  - For each Wikipedia article:
    - Rank each term by TF*IDF score
    - Save top 100 terms per article
- Saved 90 Wikipedia articles, one per topic in devtest and evaltest

Wikipedia Topic Focus: “Cyclone Sidr”

sidr 136.2750910626598
bangladesh 71.35448665114939
cyclone 60.890695229424374
foods 52.587400877907385
blankets 50.85101156044608
kmh 48.60582997871087
emergency 35.92842265916994
assistance 32.16653898594446
imd 30.974027355630056
disaster 30.446902476798044
response 29.685729189784887
taka 29.16349798722652

shelters 27.474138263315425
tents 26.706574232075003
affected 25.488629604526082
water 25.41721212110711
crescent 25.369765879728305
winds 25.228450869980467
diseases 24.264752373215675
areas 23.308651699028378
reported 22.086546133126927
cyclonic 21.92896081883586
relief 21.607802299441985
medicine 21.444739300894486
etc...

Wikipedia Topic Focus, continued

- What do we do with “cheat sheet” from each Wikipedia article?
  - Load top 100 terms per topic into summarization module
  - Tokenize each candidate sentence
  - Compare these tokens with top 100 terms from Wikipedia article
  - Use embeddings from pre-trained CBOW model
    - If similarity >= 0.75, add “bonus point” to wikiScore
    - Multiply final wikiScore by weight
    - Weight of 200 best

Optimization

- Tuned parameters to devtest ROUGE scores
- Best parameters:
    - `--size 600`
    - `--topN 60`
    - `--corpus wikipedia`
    - `--wikiScores wikipediaScores50000.json`
    - `--wikiWeight 200`
    - `--wikiIDF wikipediaIDF50000.json`
    - `--wikiCBOW wikipediaCBOW50000mincount2`

Scores - Devtest Improvements

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
D3 Scores	0.25363	0.07330	0.02577	0.01001
With Wikipedia	0.26499	0.07566	0.02768	0.01161
With Regex/POS Compression	0.28582	0.08174	0.03052	0.01323
With Parser Compression	0.26200	0.07443	0.02559	0.00994

- Decided to comment out parser compression.

Scores - Devtest & Wikipedia

Wikipedia background	Wikipedia topic focus	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
Yes	Yes	0.28582	0.08174	0.03052	0.01323
No	Yes	0.27414	0.07619	0.02701	0.00985
Yes	No	0.26389	0.06711	0.01866	0.00636
No	No	0.26674	0.06858	0.02038	0.00686

- (used Reuters from NLTK as alternate background corpus)
- (wikiWeight = 0 when testing without Wikipedia topic focus)

Scores - Compression

	Compression	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
Devtest	yes	0.28582	0.08174	0.03052	0.01323
Devtest	no	0.26329	0.07298	0.02453	0.00882
Evaltest	yes	0.32139	0.09917	0.03825	0.01826
Evaltest	no	0.29412	0.08364	0.02887	0.01247

- Compression definitely helps!

Scores - Devtest vs. Evaltest

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
Devtest	0.28582	0.08174	0.03052	0.01323
Evaltest	0.32139	0.09918	0.03795	0.01796

Summary - “China Water Shortage”

- China is among the driest countries in the world and 400 out of 600 Chinese cities suffer from water shortages for domestic and industrial uses.
- China faces a severe water shortage especially in the northern part of the country.
- The reduced water flow is affecting the river capacity to dilute pollutants.
- The Three Gorges Dam in central China Hubei Province has opened its floodgates to ease the severe water shortages along the Yangtze River.
- China central and western regions will suffer an annual water shortage of about 20 billion cubic meters from 2010 to 2030.