Extractive Summarization: MMR

- Given some articles and a length budget of *k* words, pick some sentences of total length <= *k* and make a summary
- Pick important yet diverse content: maximum marginal relevance (MMR)

While summary is < *k* words

Calculate $\text{MMR} = \text{argmax}_{D_i \in R \setminus S} \left[\lambda(\text{Sim}_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} \text{Sim}_2(D_i, D_j)) \right]$

"max over all sentences not yet in the summary"

"make this sentence similar to a query"

"make this sentence maximally different from all others added so far"

Add highest MMR sentence that doesn't overflow length

Carbonell and Goldstein (1998)

Centroid

- Represent the documents and each sentences as bag-of-words with TF-IDF weighting

While summary is < *k* words

Calculate score(sentence) = cosine(sent-vec, doc-vec)

Discard all sentences whose similarity with some sentence already in the summary is too high

Add the best remaining sentence that won't overflow the summary

Erkan and Radev (2004)

Summarization

- Count number of *documents* each bigram occurs in to measure importance

score(*massive earthquake*) = 3

score(*six killed*) = 2

score(*magnitude 7.3*) = 2

score(*Iraqi capital*) = 1

- Find summary that maximizes the score of bigrams it covers

- ILP formulation: *c* and *s* are indicator variables indexed over bigrams ("concepts") and sentences, respectively

$$\text{Maximize: } \sum_i w_i c_i \qquad s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\text{Subject to: } \sum_j l_j s_j \leq L \qquad \sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

"set c_i to 1 iff some sentence that contains it is included"

sum of included sentences' lengths can't exceed L

Gillick and Favre (2009)

Evaluation: ROUGE

- ROUGE-n: n-gram precision/recall/F1 of summary w.r.t. gold standard
- ROUGE-2 correlates somewhat well with human judgments for multi-document summarization tasks

~~*A*~~ *massive earthquake* ~~*of*~~ *magnitude 7.3 struck Iraq* ~~*on*~~ *Sunday* — prediction

~~*An*~~ *earthquake* ~~*was*~~ *detected* ~~*in*~~ *Iraq* ~~*on*~~ *Sunday* — reference

ROUGE 2 recall = 1 correct bigram (Iraq, Sunday) / 4 reference bigrams

ROUGE 2 precision = 1 correct bigram (Iraq, Sunday) / 6 predicted bigrams

- Many hyperparameters: stemming, remove stopwords, etc.
- Historically: ROUGE recall @ k {words, characters}. Now: ROUGE F1

Lin (2004)

Results

Model	R-1	R-2	R-4
Centroid	36.03	7.89	1.20
LexRank	35.49	7.42	0.81
KLSum	37.63	8.50	1.26
CLASSY04	37.23	8.89	1.46
ICSI	38.02	**9.72**	**1.72**
Submodular	38.62	9.19	1.34
DPP	**39.41**	9.57	1.56
RegSum	38.23	9.71	1.59
Better centroid:	38.58	**9.73**	1.53

Gillick and Favre / bigram recall

- Caveat: these techniques all work better for multi-document than single-document!

Ghalandari (2017)

Multi-document vs. Single-document

- *"a massive earthquake hit Iraq"* *"a massive earthquake struck Iraq"* — lots of redundancy to help select content in multi-document case

- When you have a lot of documents, there are more possible sentences to extract:

But eight villages were damaged in Iran and at least six people were killed and many others injured in the border town of Qasr-e Shirin in Iran, Iranian state TV said.

The quake has been felt in several Iranian cities and eight villages have been damaged.

- Multi-document summarization is easier?