

Summarization System – LING573



Group 10
Jiadong Li

D4



- ❧ Improvements in system ordering
- ❧ Improvements in content selection
- ❧ Improvement in content realization
- ❧ Issues and successes
- ❧ Related reading

Improvements in system ordering



❧ Previous approach:

❧ Content selection → information ordering → content realization.

❧ New approach:

❧ Content selection → Content realization → information ordering

Improvements in content selection



Previous solution:

- Selected the highest score until the next sentence breaks 100 words limit.
- problem: not fully used.

New solution:

- Because of the process of content realization, the sentences are shorten than the original length, so I will keep search for sentences to fit in the summary.
- Problem: some low score sentence are selected at the end to fulfill the words limit.
- solutions: only select from the top 20 scored sentence.

Improvements in content realization



- ✧ UMD Approach #1 Trimmer based on POS tags
 - ✧ Split the sentence by any punctuation.
 - ✧ Remove shorter phrase (≤ 2) at the beginning.
 - ✧ Remove temporal expressions
 - ✧ Remove news source
 - ✧ Remove phrase in between two '--'.
 - ✧ Remove unnecessary punctuation.
 - ✧ Remove complementizer that.
 - ✧ Remove complement phrase in between two punctuation (normally comma and period) and After a NN tag and before a VB/W tag.
 - ✧ Remove some determiners before Capitalized Proper Noun.
 - ✧ Remove some adv.
 - ✧ Remove PPs that do not contain Nes (not fully implemented yet).
- ✧ Kill unnecessary space
 - ✧ `new_sentence = new_sentence.replace(" 's", "'s")`
`new_sentence = new_sentence.replace(" 're", "'re")`
`new_sentence = new_sentence.replace(" n't", "n't")...`

Improvements in content realization



❧ Problem:

- ❧ After content realization, redundancy becomes more often.

❧ Solution:

- ❧ In the content selection step, if current sentence is quite similar to one of previous selected sentence, then skip it.
 - ❧ Similarity :if-idf comparison or word co-appearance ratio

Results



D2

	R(%)	P(%)	F(%)
ROUGE-1	17.026	24.534	19.931
ROUGE-2	4.88	7.093	5.73
ROUGE-3	1.762	2.639	2.092
ROUGE-4	0.655	1.011	0.788

D3

	R(%)	P(%)	F(%)
ROUGE-1	22.887	28.247	25.085
ROUGE-2	6.366	7.789	6.954
ROUGE-3	2.158	2.658	2.363
ROUGE-4	0.895	1.115	0.985

D4

	R(%)	P(%)	F(%)
ROUGE-1	24.605	26.078	25.156
ROUGE-2	6.611	6.936	6.751
ROUGE-3	2.072	2.159	2.108
ROUGE-4	0.786	0.812	0.796

Data used: ../devtest/GuidedSumm10_test_topics.xml
ID from D1001A to D1046H

Results



devtest

	R(%)	P(%)	F(%)
ROUGE-1	22.887	28.247	25.085
ROUGE-2	6.366	7.789	6.954
ROUGE-3	2.158	2.658	2.363
ROUGE-4	0.895	1.115	0.985

evaltest

	R(%)	P(%)	F(%)
ROUGE-1	27.908	29.233	28.463
ROUGE-2	7.452	7.783	7.593
ROUGE-3	2.445	2.532	2.482
ROUGE-4	1.039	1.056	1.045

Data used: ../devtest/GuidedSumm10_test_topics.xml
ID from D1001A to D1046H

Data used: ../evaltest/GuidedSumm11_test_topics.xml
ID from D1101A to D1144H

Issues



❧ Named Entity Redundancy

- ❧ I have tried the Stanford Named Entity Recognizer (NER) and replace them with supervised trained classifier result.
- ❧ Have some progress but not fully implemented.

Related reading



- ❧ Multi-candidate reduction: Sentence compression as a tool for document summarization tasks
- ❧ <https://www.sciencedirect.com/science/article/abs/pii/S0306457307000295>

Thank you

