*h*MDS: The Next Step for Multi-Document Summarization

TECHNISCHE UNIVERSITÄT DARMSTADT
AIPHES
Markus Zopf, Maxime Peyrard, Judith Eckle-Kohler | Research Training Group AIPHES

*h*MDS in a nutshell

- novel heterogeneous multi-document summarization corpus
- poses new challenges for summarization systems
- novel construction approach makes corpus construction easy
- get more information at `https://github.com/AIPHES/hMDS`

Traditional Corpus Construction

1. search source documents by topic
2. identify important information in source documents
3. write a proper summary

problems:
- for high text quality, professional writers are require
- exponential complexity to read all documents
- domain knowledge required to estimate importance

→ **time-consuming + expensive**

executed e.g. for DUC 2004 / TAC 2008
→ very homogeneous corpora: only English newswire source documents, always 100 word summaries, etc.

Reversing Traditional Corpus Construction

1. select a summary
2. mark important information
3. retrieve source documents by information nuggets

advantages:
- high-quality summary already available
  e.g. Wikipedia featured articles: well-written, comprehensive, well-researched, stable
  → no text to write → no professional writer needed
- importance of information already assessed
  → no topic knowledge needed
- executable without human interaction?

result of corpus creation:
- heterogeneous, multi-genre source documents
- variable length source documents & summaries
- noisy source documents

Dimensions of Heterogeneity

Genres & Textual Heterogeneity

Lots of different text genres

microblog, encyclopedic short, organization, scientific, dialogues, education, article, encyclopedic long, forum post, social media

Textual Heterogeneity

based on Jensen-Shannon (JS) divergence of word distributions

$$\mathrm{TH}_{\mathrm{JS}}(\text{topic}) = \frac{1}{|\text{topic}|} * \sum_{\text{doc} \in \text{topic}} \mathrm{JS}\left(P_{\text{doc}}, P_{\text{topic} \setminus \text{doc}}\right)$$

	*h*MDS	DUC 04	TAC 08
Avg. $\mathrm{TH}_{\mathrm{JS}}$	0.3815	0.3019	0.3188

Observation:
- different text genres introduce high textual heterogeneity

Document Lengths

Summaries

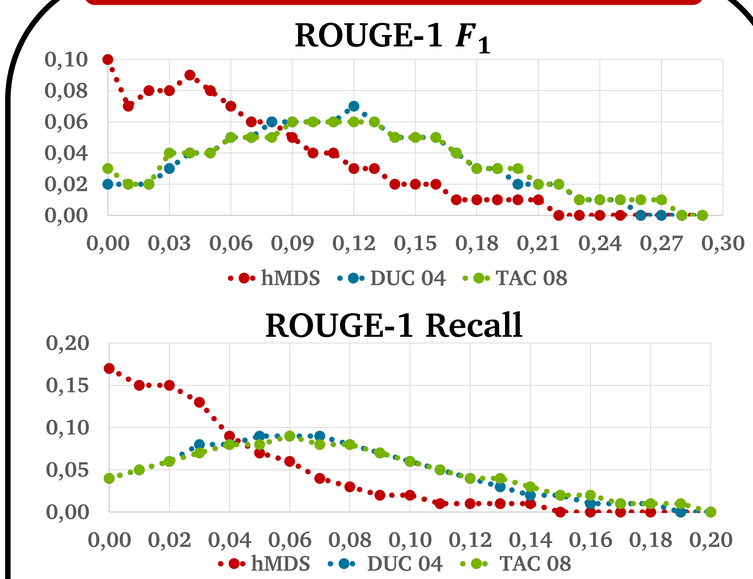
Corpus	Avg. Length (Words)	Relativ SD
*h*MDS	245.55 ± 132.94	0.54
DUC 04	118.11 ± 6.38	0.05
TAC 08	109.33 ± 7.01	0.06

Source Documents

Corpus	Avg. Length (Words)	Relativ SD
*h*MDS	1863.59 ± 3928.91	2.11
DUC 04	672.14 ± 6.38	0.75
TAC 08	589.20 ± 7.01	0.82

Observation:
- Documents in *h*MDS longer
- *h*MDS higher standard deviation
  - also in summaries!

Distribution of ROUGE Scores



Observation:
- DUC and TAC very similar
- *h*MDS much more poor sentences/noise
  → harder to summarize

Summarization Experiments

Corpus	*h*MDS-M	*h*MDS-A	*h*MDS-V	DUC 04	TAC 08
Optimal	*0.4960*	*0.4845*	*0.5018*	*0.1876*	*0.2540*
Random	0.0732	0.0594	0.0450	0.0435	0.0458
Lead	0.1237	0.0318	0.0018	0.0766	0.0765
LexRank	0.1273	0.1192	0.0797	0.0715	0.0773
ICSI	**0.2293**	**0.2267**	**0.2082**	**0.0900**	**0.1107**
LSA	0.0689	0.0652	0.0603	0.0430	0.0696
TF-IDF	0.0939	0.0805	0.0766	0.0657	0.0572

ROUGE-2 scores

Conclusions

- optimal scores much higher in *h*MDS
- ICSI performs best
- reference systems better suited to summarize classical datasets
- **larger gap to optimum: go for it!**
- LEAD (usually strong) performs poorly compared to reference systems
- performance decreases if more noise is present (M → A → V)

Challenge accepted!

get more information at
`github.com/AIPHES/hMDS`

current work on *h*MDS:
1. adding German topics
2. automating construction approach

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1

www.aiphes.tu-darmstadt.de