# auto-hMDS: Automatic Construction of a Large Heterogeneous Multilingual Multi-Document Summarization Corpus

Markus Zopf | Research Training Group AIPHES, TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

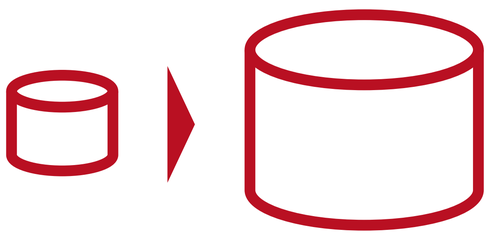
## auto-hMDS in a nutshell

- lack of large multi-document summarization corpora limits training and evaluation of machine learning models
- large corpora are expensive to create
- automatically create large, multilingual MDS corpora
- retrieve auto-hMDS corpus: `github.com/AIPHES/auto-hMDS`

## Problem

### Machine learning requires large datasets

**image classification 2014:**
GoogLeNet trained on 1.5 million images



**machine translation 2017:**
DeepL trained on over 1 billion translations

**single-document summarization:**
CNN/DailyMail corpus → abstractive summarization

### Reliable evaluation requires large datasets

- evaluation with ROUGE is noisy
- ROUGE preference prediction correct in approx. 65%
- requires more data points to be reliable

A > B ?

### Large MDS corpora are not available

Available MDS datasets are
- small: DUC/TAC contain only up to 50 topics
- written only in English
- contain only newswire topics + documents

## Solution

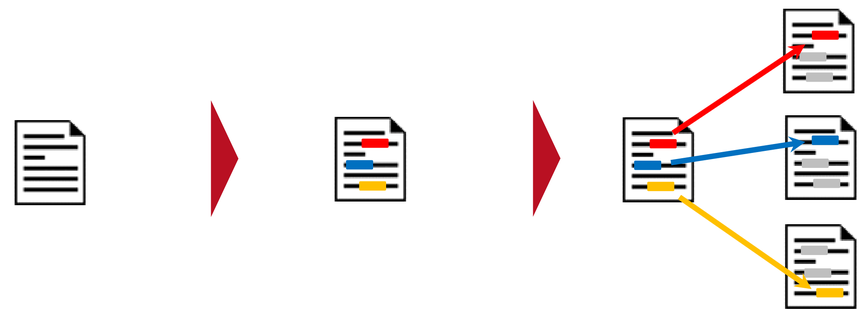
### Traditional Corpus Construction

1 search source documents for topic
2 identify important information
3 write proper summaries

1. finding/reading source documents is time consuming
2. importance estimation requires domain knowledge
3. professional writers required to achieve high quality

→ **time-consuming, complex and expensive**
→ **leads to small corpora**

### Reversing Traditional Corpus Construction



1 select existing summary
2 mark information nuggets
3 retrieve source documents

1. no text to write → cheap
2. importance of information already assessed
3. summaries = common opinion

prior work performed 2 and 3 manually to build hMDS
→ **created hMDS corpus still rather small**

**perform 2 and 3 automatically:**

**search one source document for every sentence in the summary**

## Analysis

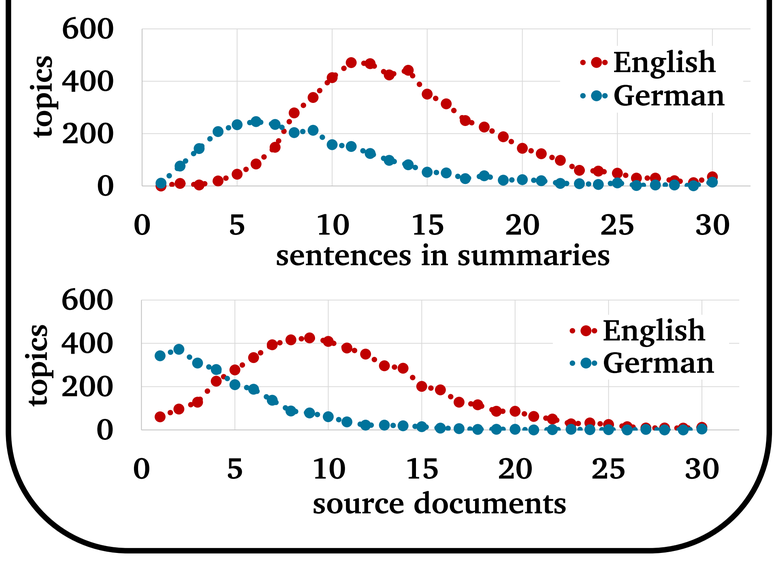
### Corpus Size

| Corpus | Topics | Source Documents |
|---|---|---|
| DUC 2004 | 50 | 500 |
| TAC 2009 | 44 | 440 |
| hMDS | 91 | 1,265 |
| auto-hMDS | **7,316** | **64,744** |

| Corpus | Sentences | | Tokens | |
|---|---|---|---|---|
| | source | sum | source | sum |
| DUC 2004 | 26.28 | 6.61 | 672.15 | 118.12 |
| TAC 2009 | 24.58 | 6.16 | 633.89 | 110.15 |
| hMDS | 268.15 | 9.05 | 2972.12 | 245.52 |
| auto-hMDS | **271.36** | **12.54** | **5862.51** | **312.42** |

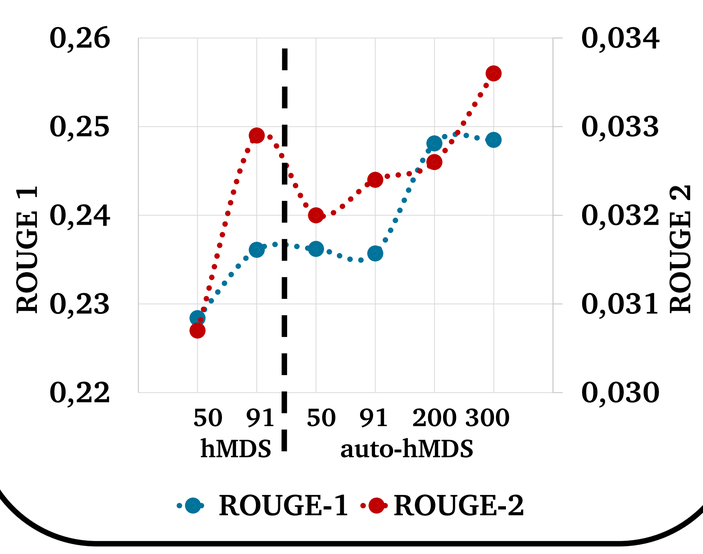
### Corpus Heterogeneity

- **broad variety of different topics**
- **topics in English and German**




### Training Data

- training on (auto-)hMDS
- evaluation on TAC 2009




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