Beyond Generic Summarization: A Multi-faceted Hierarchical Summarization Corpus of Large Heterogeneous Data

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES
ADAPTIVE PREPARATION OF INFORMATION FROM HETEROGENEOUS SOURCES

Christopher Tauchmann, Thomas Arnold, Andreas Hanselowski, Christian M. Meyer, Margot Mieskes
Research Training Group AIPHES, https://www.aiphes.tu-darmstadt.de

Summary

- New approach for creating hierarchical summarization corpora
- First, extract relevant content using crowdsourcing,
- Second, ask expert annotators to order relevant information hierarchically

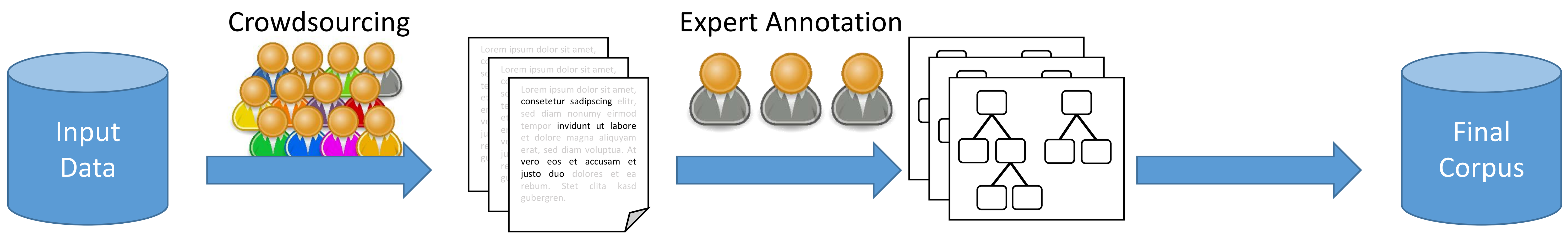
Motivation

- Automatic summarization so far focused on small datasets
- Automatic systems could analyze documents from huge document collections

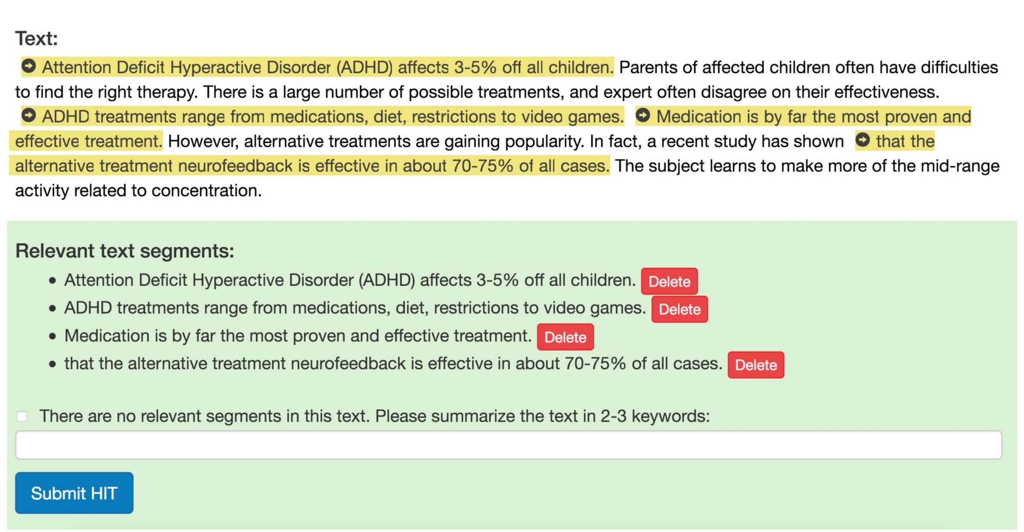
Results

- Highly heterogeneous corpus
- Crowdsourced information nuggets
- Tree structures covering the specific facets discussed in a document collection
- Can be used to develop and evaluate hierarchical summarization systems

Workflow



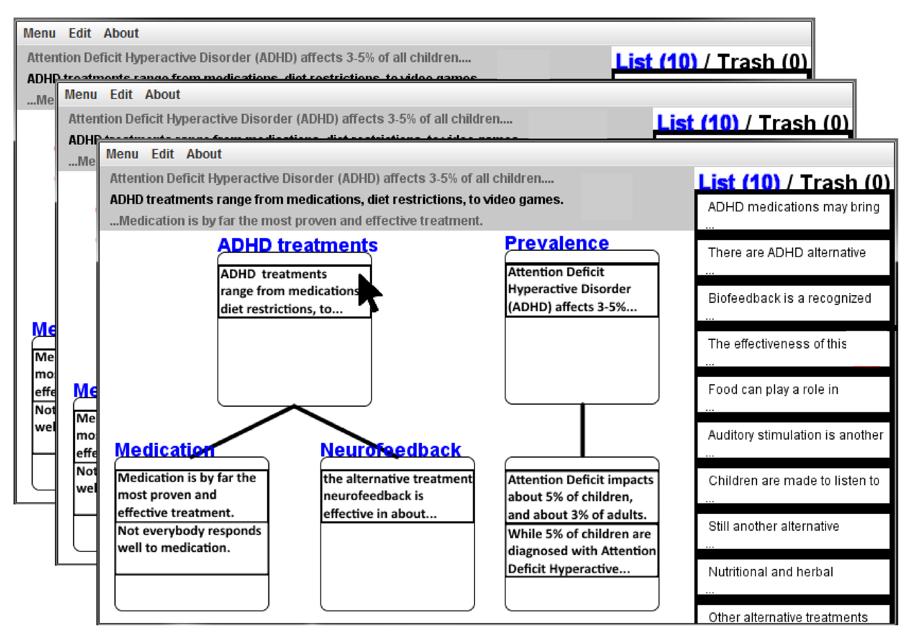
Content Selection



HITs on Amazon Mechanical Turk

- Information Nuggets Include all facts, opinions, hypotheses/statements and claims to include in a summary on given topic
- Select nuggets marked by 3+ workers

Hierarchical Ordering

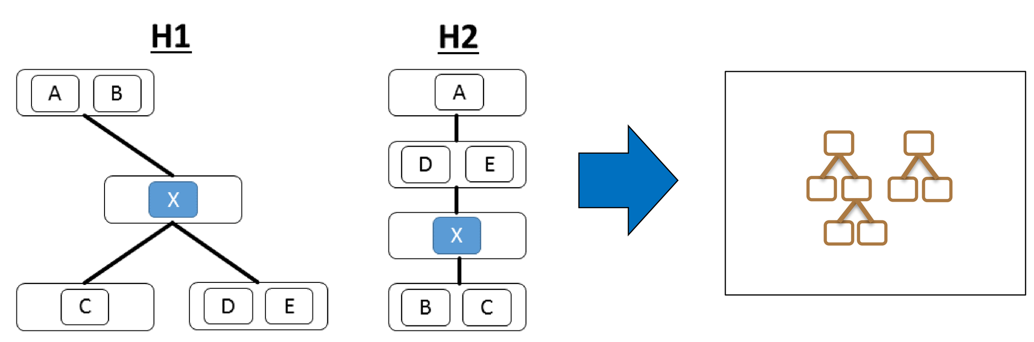


Hierarchy Annotation Tool

- Three expert annotators create hierarchy
- Hierarchies have multiple facet trees and no shared root node

Structural Analysis / Gold Standard

$$HO(H_1, H_2) = a \cdot TO(H_1, H_2) + b \cdot SupO(H_1, H_2) + c \cdot SubO(H_1, H_2)$$



Hierarchy Overlap (HO) → Gold Standard

- Compute Gold Standard with a newly developed measure
- Consecutively add each information to empty nugget hierarchy
- Successively remove each information nugget and insert it again at best possible position

https://github.com/AIPHES/HierarchicalSummarization

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