

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



TECHNISCHE
UNIVERSITÄT
DARMSTADT



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

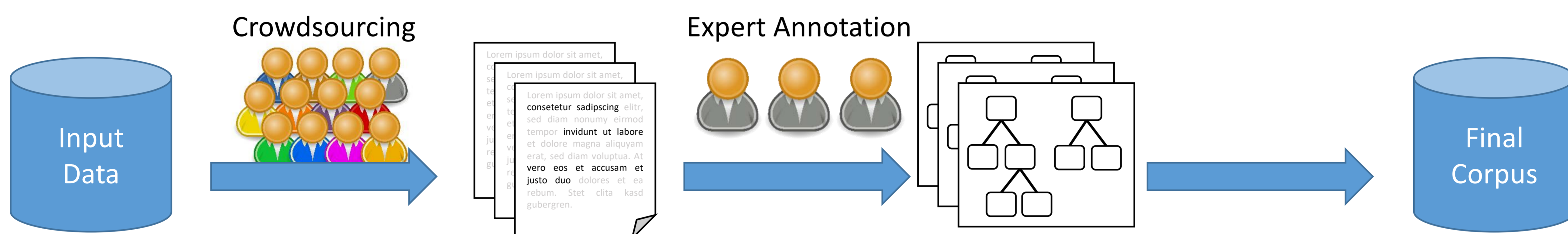
Motivation

- 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

Text:
 ● Attention Deficit Hyperactive Disorder (ADHD) affects 3-5% off all children. Parents of affected children often have difficulties to find the right therapy. There is a large number of possible treatments, and expert often disagree on their effectiveness.
 ● ADHD treatments range from medications, diet, restrictions to video games. ● Medication is by far the most proven and effective treatment. However, alternative treatments are gaining popularity. In fact, a recent study has shown ● that this alternative treatment neurofeedback is effective in about 70-75% of all cases. The subject learns to make more of the mid-range activity related to concentration.

Relevant text segments:

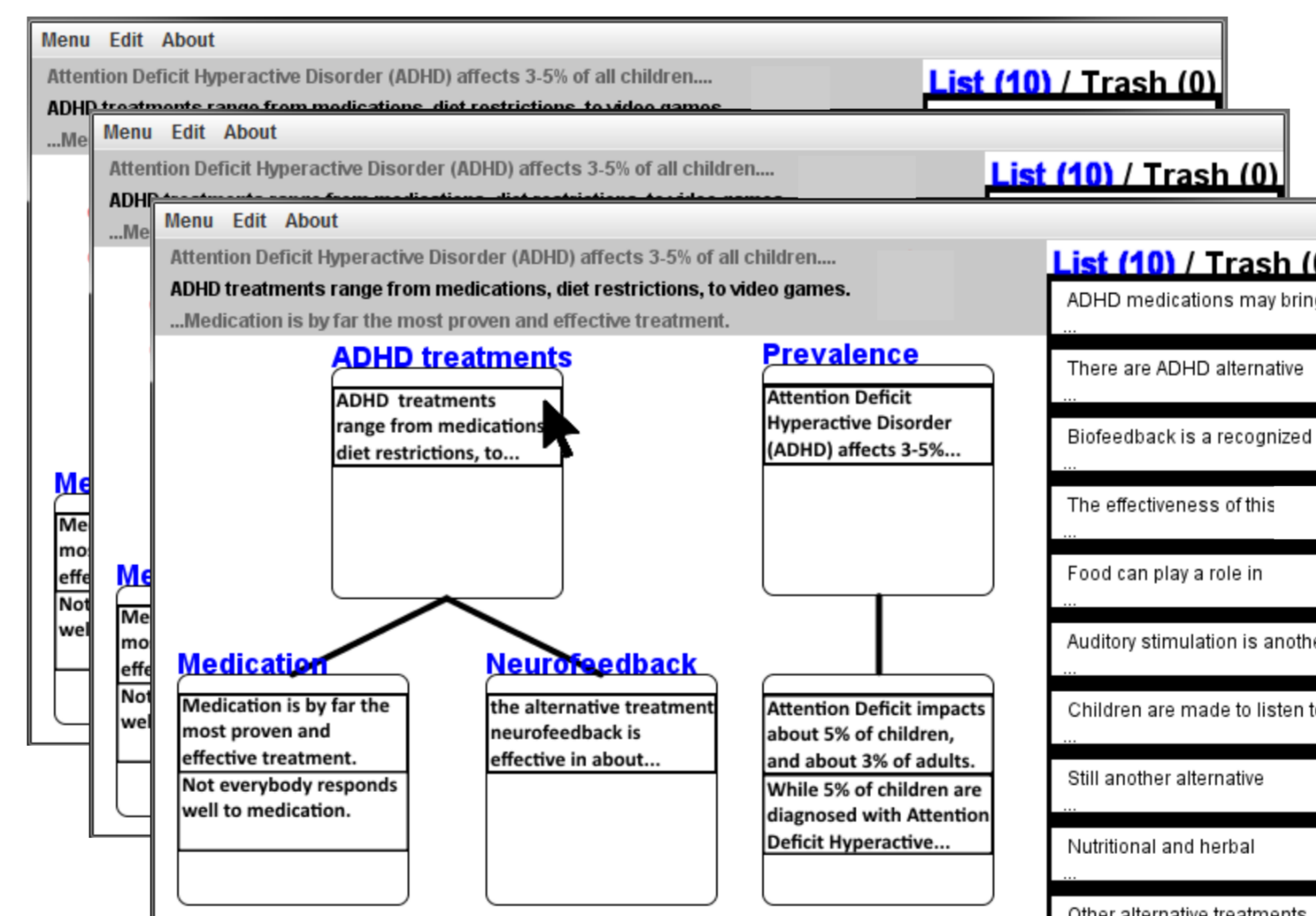
- Attention Deficit Hyperactive Disorder (ADHD) affects 3-5% off all children. Delete
- ADHD treatments range from medications, diet, restrictions to video games. Delete
- Medication is by far the most proven and effective treatment. Delete
- that the alternative treatment neurofeedback is effective in about 70-75% of all cases. Delete

There are no relevant segments in this text. Please summarize the text in 2-3 keywords:

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

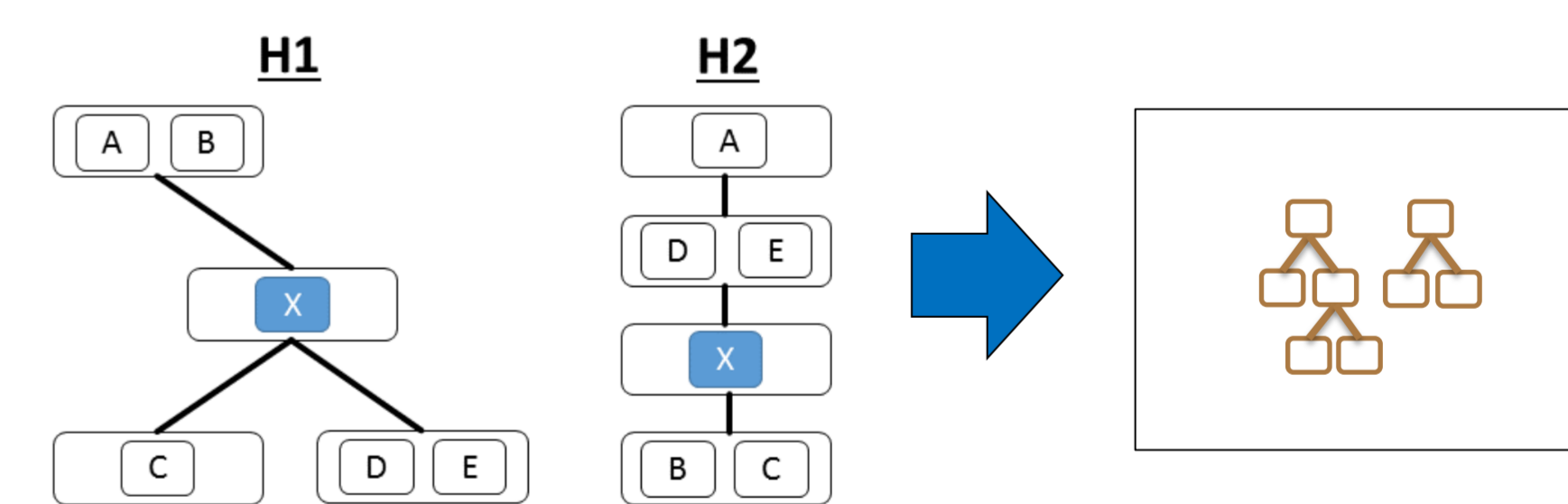


Hierarchy Annotation Tool

- 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>