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

Christopher Tauchmann*, Thomas Arnold*, Andreas Hanselowski*, Christian M. Meyer* and Margot Mieskes‡
Research Training Group AIPHES
*Technische Universität Darmstadt; ‡Hochschule Darmstadt
https://www.aiphes.tu-darmstadt.de

Abstract

Automatic summarization has so far focused on datasets of ten to twenty rather short documents, typically news articles. But automatic systems could in theory analyze hundreds of documents from a wide range of sources and provide an overview to the interested reader. Such a summary would ideally present the most general issues of a given topic and allow for more in-depth information on specific aspects within said topic. In this paper, we present a new approach for creating hierarchical summarization corpora from large, heterogeneous document collections. We first extract relevant content using crowdsourcing and then ask trained annotators to order the relevant information hierarchically. This yields tree structures covering the specific facets discussed in a document collection. Our resulting corpus is freely available and can be used to develop and evaluate hierarchical summarization systems.

Keywords: hierarchical summarization, large corpora, heterogeneous sources, crowdsourcing, aspect-oriented summarization

1. Introduction

Automatically created summaries are most useful if they allow readers to save time when reading long and/or many documents from a large number of sources. However, many state-of-the-art approaches in automatic multi-document summarization (MDS) are still evaluated on small clusters of ten to twenty short articles. The most prominent document collections from the DUC and TAC conferences have, for example, only about 6,700 (DUC '04) and 17,400 (DUC '06) tokens per topic cluster.[1] This evaluation setup does not cover the full potential of automatic summarization, which *could* easily aggregate collections of over hundred documents with more than 100,000 tokens.

In some respects, the current setup is not even very realistic, as the vast majority of the available datasets cover only newswire text about a single event or entity (Nenkova, 2005). Given the large amount of redundancy in this text type, a human reader could read only one or two of the source documents and quickly skim over the remaining ones to get a good overview of the article's main event or entity – albeit update summaries would be helpful in this situation. Even more recent work in social media and real-time summarization is based on high-redundancy text (Chua and Asur, 2013; Lin et al., 2016). In large heterogeneous document collections, there are important facts and arguments that appear only in few of the available documents and are therefore missed by generic summary strategies and absent from both automatic and reference summaries.

With increasing volume, velocity, and variety of the source documents, it gets, however, extremely difficult to construct suitable evaluation corpora. Assuming a reading speed of 228 ± 30 words per minute for English (Trauzettel-Klosinski and Dietz, 2012), it already takes more than seven hours (excluding breaks) to read a document collection with 100,000 words. It is hardly possible for an individual annotator to stay equally concentrated for that many hours. This yields a bias in the resulting summary, as the annotators will gradually shift their notion of what is important – especially in heterogeneous low-redundancy texts where frequency of occurrence is not a good indicator for importance. Although query-focused or aspect-oriented summaries yield a frequency-agnostic notion of importance, the resulting summarization corpora cover only a small fraction of the collection's content, which makes the annotation less cost-efficient. Corpora covering only a few narrow queries also lack the general overview of the large variety of facets typically discussed in broad and large collections.

In this work, we propose a novel approach to create summarization corpora for large document collections by structuring the important information hierarchically. We particularly focus on controversial topics from the educational domain, such as *alternative ADHD treatments*. This topic also serves as a running example throughout the paper, as it may be viewed from many different facets (or points of view), including ADHD prevalence, risk groups, diagnosis, nutrition treatment, herbal treatment, hypnosis, and music therapy. We would expect this kind of information in a generic summary about the topic. However, each facet should also branch off and discuss the most important symptoms for affirming or excluding a diagnosis in one branch, as well as different procedures, their advantages and disadvantages, and evidence for their effectiveness in other treatment-specific branches. A hierarchical structure of this and similarly complex topics therefore covers general information about the topic as well as detailed information on each facet discussed in the document collection. Methods for automatically creating such hierarchical summaries are highly relevant to complex information seeking processes that assist users in gaining an overview *and* diving into specific facets of a controversial topic. However, we require new hierarchical summarization corpora in or-

[1] http://duc.nist.gov, http://tac.nist.gov

der to research and evaluate automatic systems. Our approach is suitable to create such corpora for large, heterogeneous datasets of over 100,000 tokens spanning multiple genres (e.g., scientific articles, blogs, forum posts).

Our key idea is to first collect the most relevant information independent of the actual use for the summary and then identify redundancy, granularity, and facet by organizing the collected information bottom-up into a hierarchy. Each tree of this hierarchy covers a different facet discussed in the document collection, including general definitions, specific facts, and opinions. More general information resides near the root of the tree, while more specific facts and opinions branch off to deeper tree levels grouped by topical or argumentative strand. Within the same hierarchy, we also mark redundant information by combining two information nuggets in a single tree node.

Figure 1 shows an overview of our corpus construction approach. For the first step, *content selection*, we use crowdsourcing, which allows us to process large document collections. For the second step, we rely on expert annotators and provide them with clear guidelines and a novel open-source annotation tool enabling the *hierarchical organization* of the content.

The scientific community can benefit from the proposed solution in multiple ways: Our corpus of hierarchical summaries can be used as a benchmark for automatic hierarchical summarization and information structuring methods, such as the works by Christensen et al. (2014) and Erbs et al. (2013), where there is yet almost no data available. While the hierarchical structure qualifies as a useful summary in itself, our data additionally allows us to generate textual summaries based on different parts of the hierarchy. A particular advantage of this approach is that we can summarize *all* facets discussed in a document collection by summarizing each tree of the hierarchy individually. This will save much time when creating large multi-faceted summarization corpora compared to summarizing documents for a few predefined facets, as it has been done, for example, for TAC 2010. By considering a tree's depth, we additionally gain control over the length and the level of detail of the resulting summaries.

Furthermore, we provide detailed information on our crowdsourcing setup and we publish the novel annotation tool for hierarchical summarization as open-source software in order to foster the creation of new summarization corpora.[2]

2. Related Work

Christensen et al. (2014) propose automatic hierarchical summarization, but they evaluate their system using an existing news dataset without hierarchical structure and they focus mostly on the temporal clustering of news events. Given this limited evaluation setup, we see a clear demand for new evaluation corpora that explicitly contain a hierarchical organization of the source documents' information and cover text types different from news. This will also bridge the gap between research into summarization and text structuring, such as (Erbs et al., 2013; Pembe and Güngör, 2010).

Zhang et al. (2017) discuss recursive summarization for online forums. They iteratively replace parts of the discussion with summaries, yielding a hierarchy of summaries. Our work differs in that we suggest a holistic rather than an incremental approach, which allows us to group information from discussion strands that cover related topics.

Nakano et al. (2010) focus on information credibility. They create survey reports by asking expert annotators to highlight important information in crawled web documents and describe its relation to a given topic. Based on the annotated data, they formulate summaries and investigate the impact of the annotators' information credibility descriptions on the final summary. Though they also work with large document collections, their data is not publicly available.

Falke and Gurevych (2017) recently proposed concept-map-based summarization to structure information in large document collections. Their notion of a concept map yields a generic summary that conflates all facets into a single structure of about 25 related concepts. Our work differs from that, as we organize a document collection according to the multiple facets discussed in a strict hierarchy. This enables us to induce multiple aspect-oriented summaries at varying levels of detail. Additionally, we do not rely on open information extraction, which would ignore much context and abstract from complex discourse structures, such as argumentation. Instead, we work with verbatim segments of the source texts.

Li et al. (2017) raise the issue that multi-document summarization falls short of including varying facets in the source documents. They focus on news reports and related reader comments and opinions, for which they observe that information items will not be included in a summary unless they are salient – even if the information might be interesting to readers. Li et al. (2017) also discuss comments expressing sentiments that contradict the source documents. Our proposed corpus aligns well with their work, since a hierarchy contains both salient information typically found in generic reports and opinionated and controversial statements from user comments.

Query-focused summarization (Allan et al., 2008; Baumel et al., 2016) and real-time summarization (Lin et al., 2016) are similar tasks to our work, since they aim at summarizing a specific facet discussed in a document collection or address the summarization of large amounts of data. Our hierarchical corpus construction approach yields interesting evaluation data for these tasks, since query-focused summarization systems can be trained towards multiple facets discussed in a document collection at the same time, whereas real-time summarization systems have to decide about the importance even if they do not have access to all source documents yet. Hierarchical summarization systems that generate a hierarchy similar to our manually constructed ones could yield a promising solution to this task.

So far, a lot of research in automatic summarization has been done on news documents, which has a range of shortcomings, as dicussed by Zopf et al. (2016) and Benikova et al. (2016). They argue that the spectrum of possible applications is severely limited when focusing on homogeneous

[2] GitHub repository with available data and software: https://github.com/AIPHES/HierarchicalSummarization

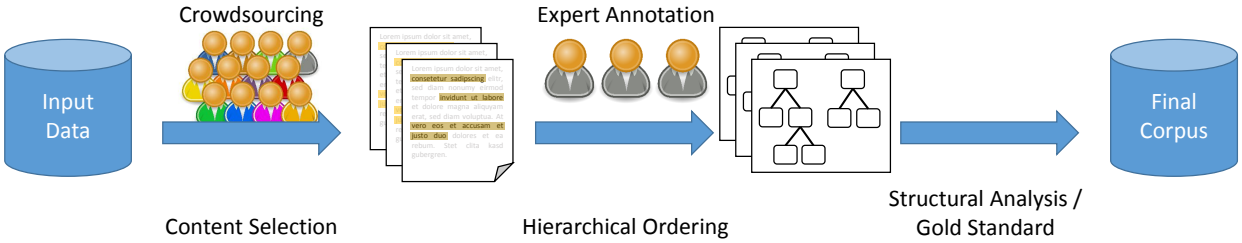

Figure 1: Overview of our corpus construction approach for hierarchically summarizing large document collections

datasets of a single text type. Both approaches propose heterogeneous summarization corpora of generic, text-based summaries, which are different from our hierarchical summaries. Nevertheless, our document collections have similar properties of incorporating heterogeneous text types.

3. Content Selection

Figure 1 shows the main steps of our corpus construction approach. In this section, we describe the content selection step, including the heterogeneous sources we use as input data, our methodology to frame the selection of important information nuggets as a crowdsourcing task, and the analysis of the resulting data.

3.1. Heterogeneous Sources

The basis for our experiment is the ClueWeb12-based focused retrieval dataset by Habernal et al. (2016). This dataset consists of 49 broad educational topic clusters with about 40–100 English documents per topic cluster. The documents are highly heterogeneous, including scientific articles, blogs, forums, personal ads, etc. Accordingly, we find both objective facts and opinionated or controversial content in this dataset. We remove duplicate sentences and documents and use only sentences that are marked relevant for a given topic in the focused retrieval dataset. This reduces the corpus from 4,820 documents with 628,026 sentences to 3,984 documents with 171,976 sentences. For our corpus, we have selected ten of those broad topic clusters. Table 1 shows the number of documents, sentences, and tokens in each topic cluster. While all topic clusters are much larger than the commonly used DUC '06 data, we sample three large (> 125,000 tokens), four medium-sized (> 50,000), and three smaller topic clusters (< 50,000). This allows us to analyze the scalability of our corpus construction approach.

3.2. Crowdsourcing and HIT Design

For the selection of important content, we use crowdsourcing. This allows us to process large document collections by breaking down the complex task into many small microtasks (Cheng et al., 2015) – so-called human intelligence tasks (HIT). Since Lloret et al. (2013) report unsatisfactory results when crowdsourcing extractive summarization, we propose a different crowdsourcing setup and further break down the summarization task into manageable microtasks by asking the crowd workers to *collect* what they consider relevant for a summary rather than to assess or rank the importance of each information at the same time.

Topic clusters	Doc.	Sent.	Tokens
Concerns about religious classes	87	7,654	210,211
School punishment policy	89	6,409	149,268
Parents of kids doing drugs	78	6,183	125,584
Children's obesity	90	3,916	90,963
Sleep problems in preschools	86	3,119	65,216
Student loans	95	2,346	54,434
Discipline in elementary school	83	2,586	53,592
Alternative ADHD treatments	57	1,475	28,281
Kids with depressions	39	1,209	21,644
Cellphone use in schools	61	902	21,384
Total	786	38,304	820,577

Table 1: Overview of our document collections and topics

We therefore generate HITs showing seven consecutive sentences from our input data at a time. In each HIT, we ask the crowd workers to mark all facts, opinions, hypotheses/statements and claims (called *information nuggets* henceforth) that they would include in a summary on the overall topic of the document collection. Our notion of information nugget is similar to previous definitions of nugget (Voorhees, 2004; Benikova et al., 2016) and semantic content unit (Nenkova et al., 2007). Workers should select only information nuggets of at least three words and a maximum length of one sentence. Each nugget should include a verb and be understandable without further context. The workers may identify multiple information nuggets within a HIT. In case they cannot find any relevant nugget, we ask them to describe the document's content to avoid spammers. Below the task description, we show two examples to illustrate the HIT. Along with the full paper, we provide a HIT template and all collected data.

Figure 2 shows a HIT for our running example. The task description is located at the top of the page. Using the *examples* button, the workers can show or hide a number of annotated examples to understand the task. Recurring workers doing multiple HITs typically do not need the examples anymore, but immediately start the annotation. They create an information nugget by clicking on its first and last word in the text. The spanned words will then be highlighted in yellow and the information nugget will be listed as a *relevant text segment*. If workers cannot find any information nuggets in a text, we ask them to summarize the text in two to three keywords. This enforces involvement and prevents workers from submitting HITs without carefully reading them.

We determine the optimal task length, payment, and num-

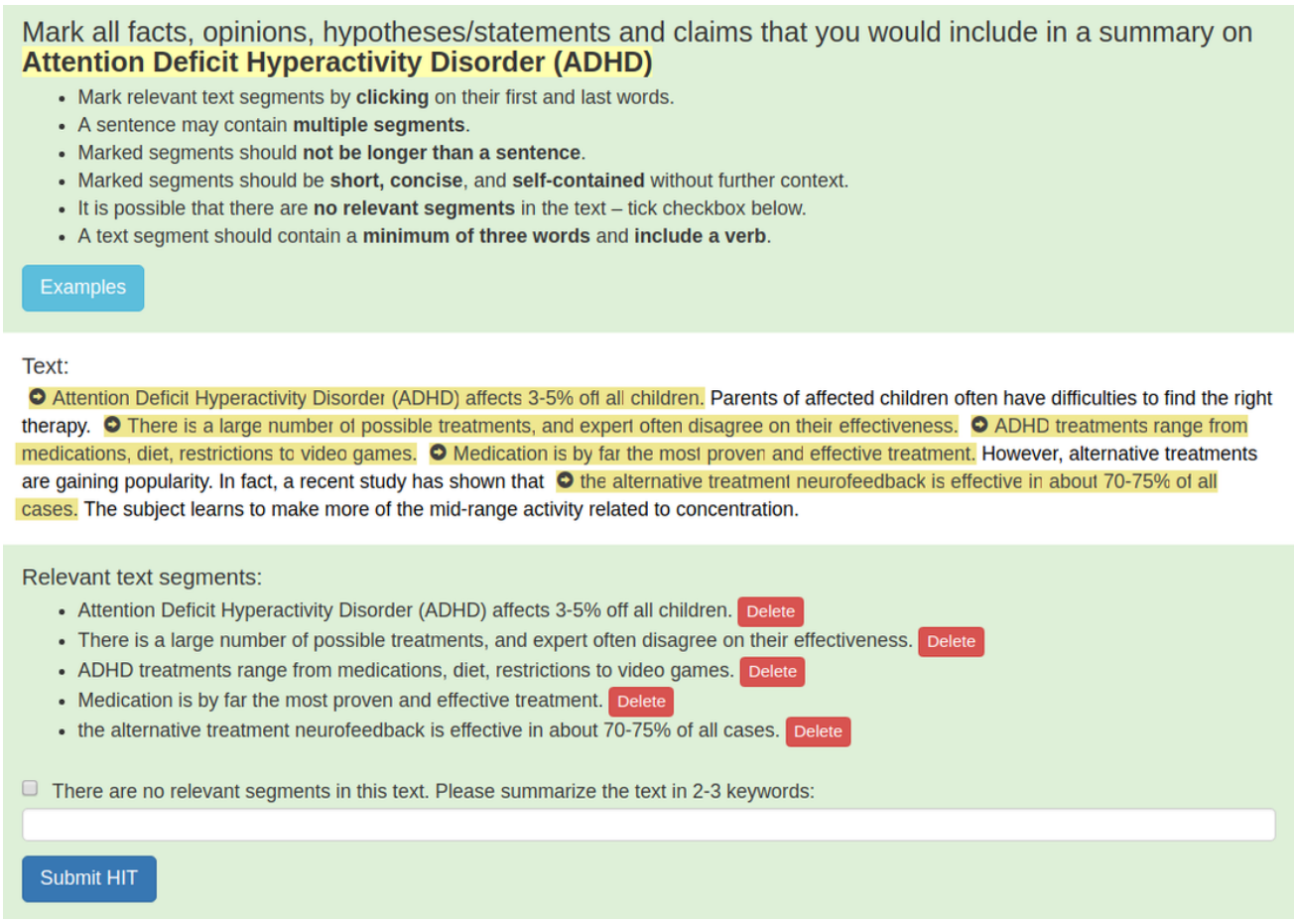

Figure 2: Screenshot of a HIT for the alternative ADHD treatments topic cluster

ber of annotators in a preliminary study. As a good trade-off between the number of HITs and the amount of work, we suggest to show short paragraphs of seven sentences in a single HIT. For each completed HIT, we pay US$ 0.07, which we find reasonable for a task of 60–90 seconds. The payment is high enough to attract reliable workers, while discouraging spammers. As quality is hard to control in a crowdsourcing setup (Bigham et al., 2015), we assign each HIT to seven workers. We select only workers with an acceptance rate of at least 98 %, we manually check annotations, reject work that does not meet our standards, and block workers where necessary.

3.3. Inter-Annotator Agreement

The crowd workers marked 68,220 information nuggets in total. Table 2 shows their inter-annotator agreement, computed using three commonly used metrics: percentage agreement A_O, Fleiss' κ (Fleiss, 1971), and Krippendorff's α_U (Krippendorff, 1995) as implemented in DKPro Agreement (Meyer et al., 2014). While A_O and κ measure agreement at the token level, α_U considers agreement between spans of selected tokens (i.e., the entire information nuggets). Both κ and α_U are chance-corrected agreement metrics (Artstein and Poesio, 2008).

The first row of Table 2 shows the scores for annotator agreement between all seven workers. The agreement is similar to previous work in summarization (Zechner, 2002; Benikova et al., 2016). In the second to fourth row, we report the agreement for the small, medium-sized, and large topic clusters individually without noticing a clear drop in annotation quality. This confirms that our crowdsourcing setup scales to large document collections.

	A_O	κ	α_U
All crowd workers	0.664	0.149	0.201
only large topic clusters	0.691	0.152	0.222
only medium topic clusters	0.634	0.127	0.189
only small topic clusters	0.666	0.170	0.186
MACE vs. Experts	0.688	0.314	0.311

Table 2: Inter-annotator agreement

To validate our results, we compare the best annotations of the seven workers according to MACE (Hovy et al., 2013) to an expert annotator, who selected information nuggets from 322 sentences. The results in the fifth row show that we reach relatively high agreement, with κ of 0.311 and α_U of 0.314. This indicates that the crowd workers selected reliable information nuggets.

3.4. Gold Standard

Most of the 68,220 information nuggets have been annotated by just a single crowd worker. To avoid singular nugget selections for the nonce, we consider only nuggets for our corpus that have been selected by at least three annotators. We remove nuggets shorter than three tokens

and merge overlapping ones. This remaining dataset has 4,983 information nuggets (7.3% of the original information nuggets), which is a manageable size for expert annotation. Within our corpus repository, we provide the source documents, the original information nuggets from Amazon Mechanical Turk, and the post-processed nuggets that serve as input for the annotation tool. The annotations are licensed under CC-BY 4.0.

4. Hierarchical Ordering

After collecting the information nuggets through a crowdsourcing approach, we structure them into hierarchies. We propose a new annotation process and a tool supporting this process. We analyze the resulting hierarchies by means of a novel evaluation metric we call *hierarchy overlap*. We finally discuss the resulting gold standard corpus of multi-faceted hierarchical summaries.

4.1. Expert Annotation and Annotation Tool

A *hierarchy* $H(V, E)$ is a forest – i.e., a directed and acyclic graph with a set of nodes V and a set of hierarchical relations $E \subseteq V \times V$. Each node $v \in V$ contains one or more information nuggets. Thus, V is a partition of the set of all information nuggets N with $\bigcup_{i \in V} v_i = N$. Each edge $(v_1, v_2) \in E$ connects more general nuggets in v_1 with more specific nuggets in v_2 discussing the same facet. There is no shared root node, so the hierarchy typically consists of multiple *facet trees*. Each facet tree contains all nuggets from one facet of the overarching topic (e.g., prevalence of ADHD), which branches off from general (e.g., overall average prevalence) to more specific information (e.g., prevalence among certain age groups or regions. To create such a hierarchy, an annotator needs to find the globally best position within the current facet trees or start a new one. The results by Lloret et al. (2013) suggest that this task cannot be broken down to a crowdsourcing setup without suffering quality problems. Therefore, we hire three expert annotators from the field of computational linguistics. This is reasonable, since the amount of data that remains after the content selection step is manageable.

To allow for an efficient annotation, we have developed a novel open-source hierarchy annotation tool with a graphical user interface. Figure 3 shows a screenshot. Input for this tool is a list of information nuggets with unique IDs and additional context from the source text, in our case the preceding and succeeding sentence.

Our tool presents a list of information nuggets that still have to be included in the hierarchy, and a working space displaying the current state of the hierarchy. Information nuggets can be added as new nodes, or into existing nodes to indicate redundant information. Alternatively, the user may structure nodes both vertically by descending salience and granularity and horizontally in new facet trees if they discuss a new facet of the overall topic. The output of the tool is the hierarchical structure in a simple XML file format.

4.2. Qualitative Analysis

For the three largest topic clusters, the annotators created hierarchies that contain 10 to 30 facet trees with an average depth of five levels. They require about six hours on average per topic cluster. One beneficial characteristic of the hierarchical structures is that different facets of controversial topics are naturally structured. Thereby, the parent node represents a specific facet and the leaf nodes different viewpoints. In the topic cluster on alternative ADHD treatments, for example, the annotators have decided to distinguish different kinds of treatments and collected claims and evidence which confirm or refute their effectiveness. Table 3 shows the number of nodes, facet trees, and average facet tree depth of all annotated hierarchies per topic.

Our qualitative analysis shows that annotators are able to structure the facets of a topic in different parts of a hierarchy. Motivated by these results, we quantify the annotators' agreement on creating the hierarchies.

4.3. Structural Analysis

To compare two hierarchies H_1 and H_2 for the same topic cluster and nugget set N, we use a modification of the *taxonomy overlap* (Maedche and Staab, 2002)

$$TO(n, H_1, H_2) = \frac{|SC(n, H_1) \cap SC(n, H_2)|}{|SC(n, H_1) \cup SC(n, H_2)|}$$

where $SC(n, H)$ is the set of all nuggets contained in sub- or supernodes (the semantic cotopy) of the node containing information nugget $n \in N$ in hierarchy H.

The averaged similarity between two hierarchies is the sum of the taxonomy overlap of all nuggets, normalized by the number of nuggets:

$$TO(H_1, H_2) = \frac{1}{|N|} \sum\nolimits_{n \in N} TO(n, H_1, H_2)$$

This metric was originally developed to measure the similarity between taxonomies and ontologies. It has been used and adapted for a variety of tasks (Euzenat and Shvaiko, 2007). However, in this metric, the order of the nodes is not important, as the metric should also compare ontologies with symmetric relations (e.g., similar-to). In our work, the relations are strictly hierarchical. Using the TO metric, a hierarchy H_1 with edges $(v_1, v_2), (v_2, v_3) \in E_1$ ("v_1 over v_2 over v_3") compared to a hierarchy H_2 with edges $(v_3, v_2), (v_2, v_1) \in E_2$ ("v_3 over v_2 over v_1") would yield a score of $TO(H_1, H_2) = 1$ (a perfect match), which contradicts our notion of a hierarchy branching from general to specific information.

Therefore, we propose our new modification called the *hierarchy overlap*

$$\begin{aligned} HO(H_1, H_2) = a \cdot TO(H_1, H_2) \\ + b \cdot SupO(H_1, H_2) + c \cdot SubO(H_1, H_2) \end{aligned}$$

which is the weighted sum of TO, the superset overlap $SupO$, and the subset overlap $SubO$ score. We compute $SupO$ and $SubO$ from taxonomy overlap TO variants that replace the full semantic cotopy SC with the nugget set of sub- *or* supernodes, respectively. Choosing the right values for the parameters a, b and c sets a trade-off between overall facet tree content and correct ordering. For our scenario, we create a small test case, explore different values

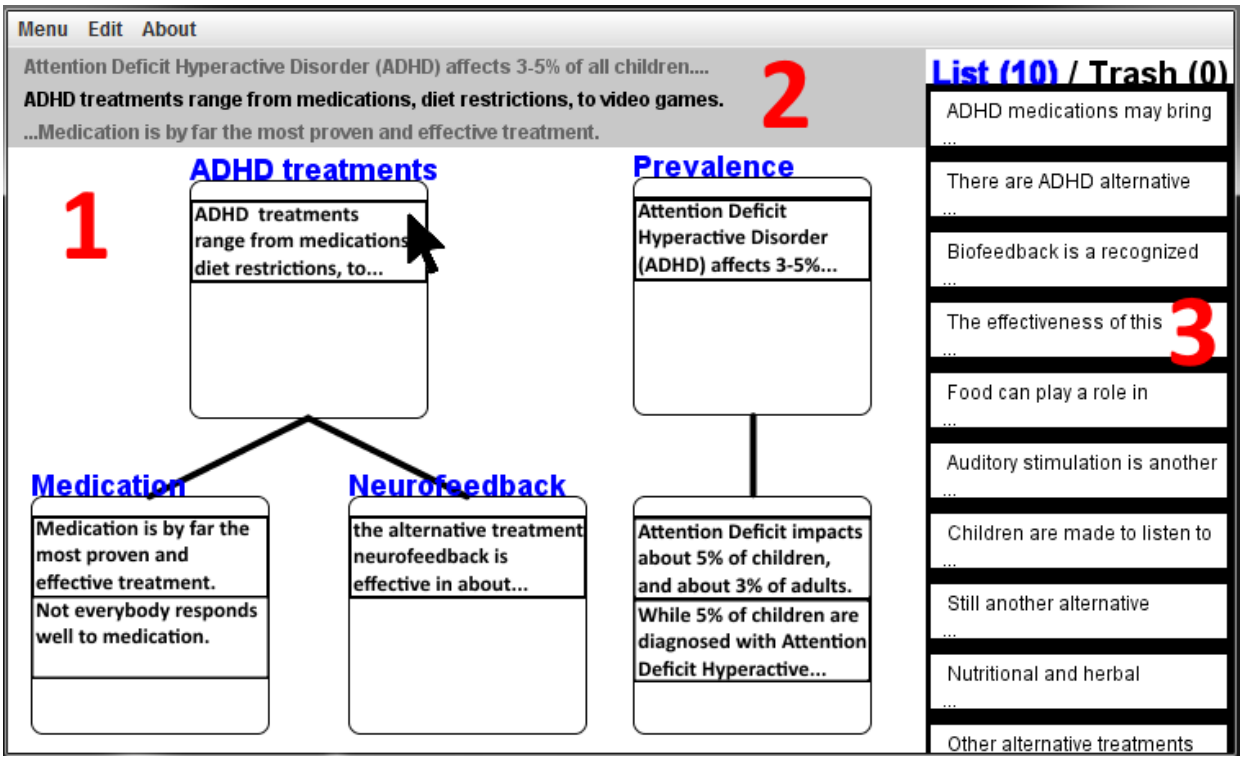

Figure 3: Screenshot of the annotation tool user interface. Area 1 is the main working space, with two annotated facet trees. Area 2 shows the full text of the hovered nugget, with preceding and succeeding sentences from the original document as context. Area 3 is a list of remaining nuggets that still have to be included in the hierarchy.

Topic	Nuggets	Nodes			Facet trees			Depth		
		A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
Concerns about religious classes	717	705	706	711	33	81	20	5.42	2.23	3.80
School punishment policy	796	704	787	747	22	29	13	5.45	2.55	6.62
Parents of kids doing drugs	1,221	1,033	1,214	1,132	31	139	10	5.35	2.06	7.50
Children's obesity	445	415	441	434	10	60	11	8.80	2.25	4.45
Sleep problems in preschools	408	401	400	390	17	56	5	7.35	2.25	8.60
Student loans	586	521	586	507	26	44	15	5.92	2.34	4.20
Discipline in elementary school	341	334	338	336	23	48	14	5.13	2.50	3.42
Alternative ADHD treatments	235	185	221	204	14	13	5	3.00	3.77	4.80
Kids with depressions	146	144	143	144	4	33	6	8.50	2.03	6.00
Cellphone use in schools	88	86	88	88	3	25	8	8.00	1.76	4.38

Table 3: Input nuggets, number of nodes, facet trees and average facet tree depth of final hierarchies (3 annotators per topic)

for the parameters and evaluate them manually. Since the partitioning of information nuggets into facet trees is our biggest priority, we use $a = 0.8$ and $b = c = 0.1$. In this case, $SupO$ and $SubO$ do not have major impact, but act as tie breakers to ensure correct information nugget order. The final HO score is still between 0 and 1.

As a simple baseline, we compute HO on randomly generated hierarchies for every topic cluster, which is between 0.09 and 0.15, depending on the topic size. In comparison, the pairwise HO of the three manually annotated hierarchies is between 0.16 and 0.28. The higher hierarchical overlap indicates that the expert annotators did agree on substantial parts of the hierarchies.

Hierarchy Overlap Example

Figure 4 shows two example hierarchies. The semantic cotopy of nugget X in hierarchy H_1 consists of all nuggets contained in sub- or supernodes of X, {A, B, C, D, E}. The semantic cotopy of nugget X in H_2 is exactly the same set. Therefore, the taxonomy overlap of nugget X in hierarchies H_1 and H_2 equals

$$\frac{|SC(X,H_1) \cap SC(X,H_2)|}{|SC(X,H_1) \cup SC(X,H_2)|} = \frac{|\{A,B,C,D,E\}|}{|\{A,B,C,D,E\}|} = 1$$

The intersection of the respective **super**sets consists of only one nugget $\{A\}$, the union has four nuggets $\{A, B, D, E\}$. The superset overlap $SupO(H_1, H_2)$ equals

$$\frac{|SupS(X,H_1) \cap SupS(X,H_2)|}{|SupS(X,H_1) \cup SupS(X,H_2)|} = \frac{|\{A\}|}{|\{A,B,D,E\}|} = \frac{1}{4}$$

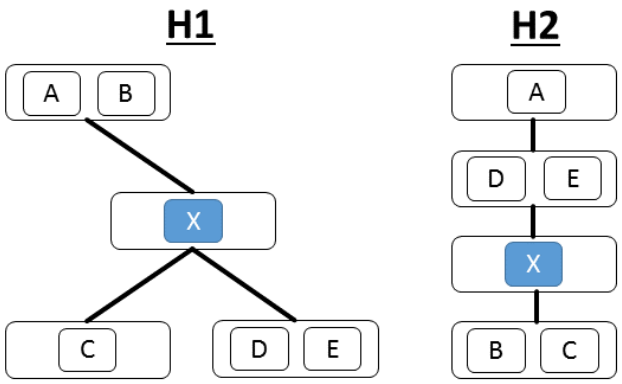

Figure 4: Hierarchy Overlap Example figure (see section 4.3. for explanations)

with the set of all nuggets $SupS(n, H)$ contained in supernodes of the node containing nugget n.
Similarly, the intersection of the **sub**sets consists of only one nugget $\{C\}$, the union has four nuggets $\{B, C, D, E\}$. The subset overlap $SubO(X, H_1, H_2)$ is $\frac{1}{4} = 0.25$. With $a = 0.8$ and $b = c = 0.1$, as proposed, the hierarchy overlap of nugget X equals

$$HO(X, H_1, H_2) = 0.8 * 1 + 0.1 * 0.25 + 0.1 * 0.25 = 0.85$$

4.4. Gold Standard

The proposed comparison metric HO enables us to create a gold standard hierarchy H_G from the three manually annotated hierarchies H_1, H_2, and H_3 for a given topic cluster. In this automatic process, we consecutively add each information nugget $n \in N$ to an empty hierarchy with a greedy strategy in order to maximize $\frac{1}{3}\sum_{i=1}^{3} HO(H_G, H_i)$. Then, we improve the resulting hierarchy with a local optimization method: We successively remove each information nugget from H_G and insert it again at the best possible position, again maximizing $\frac{1}{3}\sum_{i=1}^{3} HO(H_G, H_i)$. We repeat this process until there are no further changes. Since this local optimization can technically run into any (possibly bad) local optima, we analyze the effects of different random seeds. For one topic cluster, we perform the gold standard construction with ten differently shuffled nugget insertion orders. The normalized hierarchical overlap to the three manually annotated hierarchies varies from 0.464 to 0.496, with a mean of 0.481 and a standard deviation of 0.010. This shows that the initial position within the result space does influence the optimization result, but the effects are small. Therefore, we run each optimization with ten different random seeds and use the result with the highest $\frac{1}{3}\sum_{i=1}^{3} HO(H_G, H_i)$ as the gold standard.
In our corpus repository, we provide the Java source code of the hierarchy annotation tool, a runnable jar-file, all manually annotated hierarchies by the three annotators, and the gold standard hierarchies per topic in XML format. The software is licensed under the GNU General Public License v3.0.

5. Conclusion and Future Work

We introduced a novel approach to construct hierarchical summarization corpora, which enables us to summarize information from large document collections in a structured way. The resulting hierarchical summaries can be viewed from two perspectives: The root nodes and main branches of each tree in the hierarchy can be considered a generic summary, while each individual tree focuses on a specific facet discussed in the document collection yielding multiple aspect-oriented summaries. Our corpus can be used in a variety of problem settings within the field of automatic summarization, including table-of-contents generation, information exploration, structuring argumentative information, but also generic and query-based summarization. The logical next step is to use our corpus to train and evaluate automatic hierarchical summarization systems. We are not aware of any other dataset which can be used to evaluate all steps of such a system. Based on our annotation tool and HIT design, our approach can be easily reused by other researchers working on similar corpora for other domains or languages.

Acknowledgements

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.

6. Bibliographical References

Allan, J., Carterette, B., Aslam, J. A., Pavlu, V., Dachev, B., and Kanoulas, E. (2008). Million Query Track 2007 Overview. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC)*, pages 85–104, Gaithersburg, MD, USA.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Baumel, T., Cohen, R., and Elhadad, M. (2016). Topic Concentration in Query Focused Summarization Datasets. In *Proceedings of the 30th National Conference on Artificial Intelligence (AAAI)*, pages 2573–2579, Phoenix, AZ, USA.

Benikova, D., Mieskes, M., Meyer, C. M., and Gurevych, I. (2016). Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1039–1050, Osaka, Japan.

Bigham, J. P., Bernstein, M. S., and Adar, E. (2015). Human-computer interaction and collective intelligence. In *Handbook of Collective Intelligence*, pages 57–84. Cambridge: MIT Press.

Cheng, J., Teevan, J., Iqbal, S. T., and Bernstein, M. S. (2015). Break It Down: A Comparison of Macro- and Microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 4061–4064, Seoul, Republic of Korea.

Christensen, J., Soderland, S., Bansal, G., and Mausam. (2014). Hierarchical summarization: Scaling up multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computlational Linguistics (ACL)*, pages 902–912, Baltimore, MD, USA.

Chua, F. C. T. and Asur, S. (2013). Automatic Summarization of Events from Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 81–90, Cambridge, MA, USA.

Erbs, N., Gurevych, I., and Zesch, T. (2013). Hierarchy Identification for Automatically Generating Table-of-Contents. In *Proceedings of 9th Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 252–260, Hissar, Bulgaria.

Euzenat, J. and Shvaiko, P. (2007). *Ontology Matching*. Berlin/New York: Springer.

Falke, T. and Gurevych, I. (2017). Bringing Structure into Summaries: Crowdsourcing a Benchmark Corpus of Concept Maps. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2969–2979, Copenhagen, Denmark.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Habernal, I., Sukhareva, M., Raiber, F., Shtok, A., Kurland, O., Ronen, H., Bar-Ilan, J., and Gurevych, I. (2016). New Collection Announcement: Focused Retrieval Over the Web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 701–704, Pisa, Italy.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*, pages 1120–1130, Atlanta, GA, USA.

Krippendorff, K. (1995). On the reliability of unitizing contiguous data. *Sociological Methodology*, 25:47–76.

Li, P., Bing, L., and Lam, W. (2017). Reader-Aware Multi-Document Summarization: An Enhanced Model and The First Dataset. In *Proceedings of the EMNLP Workshop on New Frontiers in Summarization*, pages 91–99, Copenhagen, Denmark.

Lin, J., Roegiest, A., Tan, L., McCreadie, R., Voorhees, E., and Diaz, F. (2016). Overview of the TREC 2016 real-time summarization track. In *Proceedings of the 25th Text REtrieval Conference (TREC)*, Gaithersburg, MD, USA.

Lloret, E., Plaza, L., and Aker, A. (2013). Analyzing the capabilities of crowdsourcing services for text summarization. *Language Resources and Evaluation*, 47(2):337–369.

Maedche, A. and Staab, S. (2002). Measuring Similarity between Ontologies. In Asunción Gómez-Pérez et al., editors, *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web: 13th International Conference, EKAW 2002 Sigüenza, Spain, October 1–4, 2002 Proceedings*, pages 251–263. Berlin/Heidelberg: Springer.

Meyer, C. M., Mieskes, M., Stab, C., and Gurevych, I. (2014). DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING): System Demonstrations*, pages 105–109, Dublin, Ireland.

Nakano, M., Shibuki, H., Miyazaki, R., Ishioroshi, M., Kaneko, K., and Mori, T. (2010). Construction of Text Summarization Corpus for the Credibility of Information on the Web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 3125–3131, Valletta, Malta.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Trans. Speech Lang. Process.*, 4(2), May.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*, pages 1436–1441, Pittsburgh, PA, USA.

Pembe, F. and Güngör, T. (2010). A Tree Learning Approach to Web Document Sectional Hierarchy Extraction. In *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART)*, pages 447–450, Valencia, Spain.

Trauzettel-Klosinski, S. and Dietz, K. (2012). Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST. *Investigative Opthalmology & Visual Science*, 53(9):5452.

Voorhees, E. M. (2004). Overview of the TREC 2003 Question Answering Track. In *Proceedings of The Twelfth Text REtrieval Conference (TREC)*, pages 54–68, Gaithersburg, MD, USA.

Zechner, K. (2002). Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics*, 28(4):447–485.

Zhang, A. X., Verou, L., and Karger, D. (2017). Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pages 2082–2096, Portland, OR, USA.

Zopf, M., Peyrard, M., and Eckle-Kohler, J. (2016). The Next Step for Multi-Document Summarization: A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1535–1545, Osaka, Japan.