A multi-Faceted Hierarchical Summarization Corpus of Large Heterogeneous Data

Christopher Tauchmann, Thomas Arnold,
Andreas Hanselowski, Christian M. Meyer, Margot Mieskes

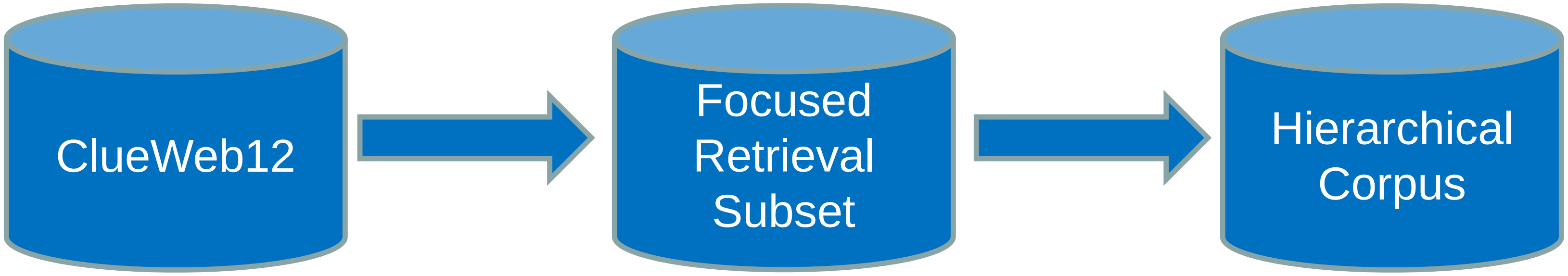
AIPHES
ADAPTIVE INFORMATIONSAUFBEREITUNG AUS HETEROGENEN QUELLEN

RUPRECHT-KARLS-
UNIVERSITÄT
HEIDELBERG

Heidelberger Institut für
Theoretische Studien

HITS

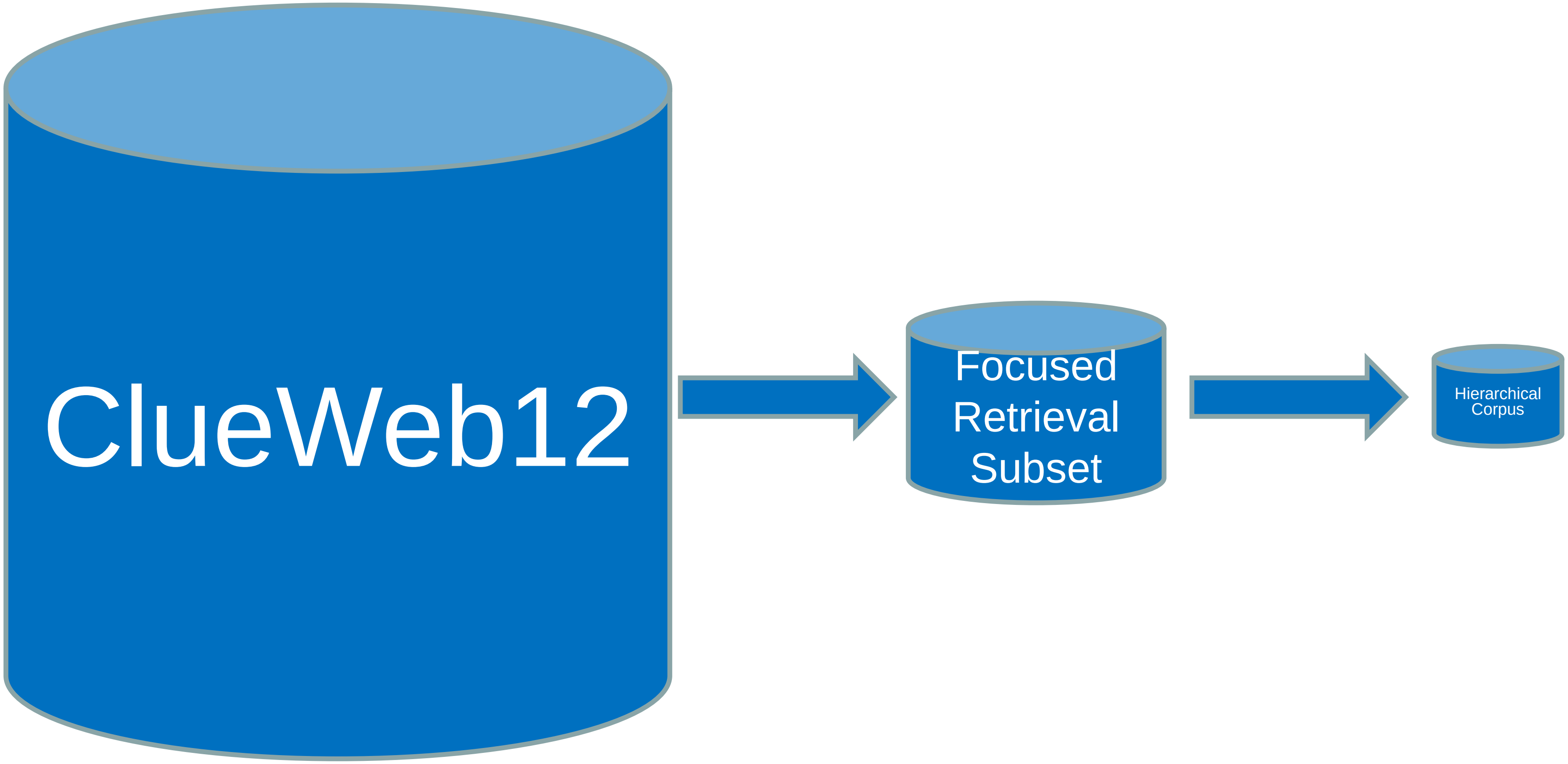
TECHNISCHE
UNIVERSITÄT
DARMSTADT



ClueWeb12

Focused Retrieval Subset

Hierarchical Corpus


ClueWeb12
Focused Retrieval Subset
Hierarchical Corpus

ClueWeb12

- **Web crawl** between February 10 and May 10, 2012 (runtime!)

Post-processing eliminates:

- Non-English pages
- Non-appropriate content
- Error code pages

- 733,019,372 English web pages
- Full data set costs 380$ (including two 3.0 TB hard disks)
- Unpacked size: 27.3 TB

Focused Retrieval Corpus
Habernal et. al.

- Filter ClueWeb12 for 49 **educational topics** (Queries)
- Examples:
  - Alternative ADHD treatments
  - Cellphone for 12 years old kids
  - Child depression

One topic cluster for each query

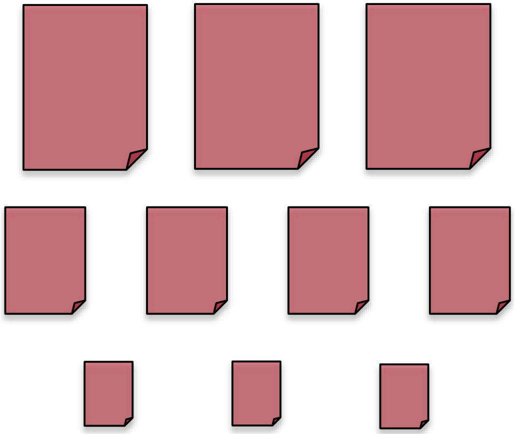
- Automatic task: Rank all documents with respect to query, take top 100
- Crowdsourcing task on Amazon Mechanical Turk:
  - Read document (or paragraph)
  - Annotate sentences: Clearly relevant, relevant in context, clearly irrelevant

Hierarchical Summarization Corpus
Tauchmann et. al. (AIPHES)

- Goals:
  - Create summarization corpus of large, heterogeneous text collection
  - Summary should contain different facets of one main topic

- Remove duplicate sentences
- Only use relevant sentences
- Result: 630,000 -> 171,976 sentences

- Choose ten topic clusters
  - Three large (> 125,000 tokens)
  - Four medium (> 50,000 tokens)
  - Three small (< 50,000 tokens)



Hierarchical Summarization Corpus

- Next step: Select important content

- Problem: Task is <u>still</u> too big for expert annotators

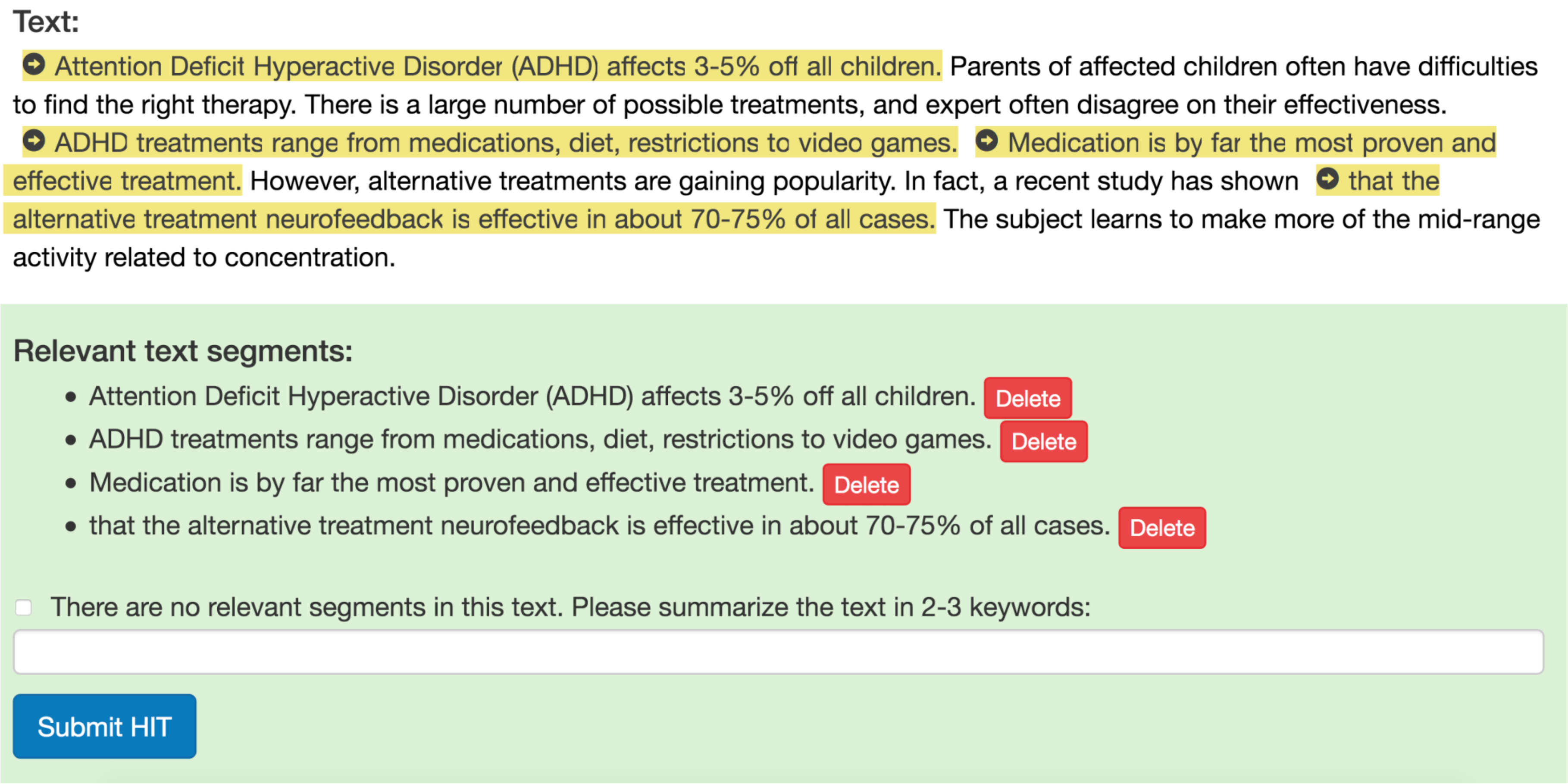
(786 documents, 38,304 sentences)

- Solution: Crowdsourcing! (Amazon Mechanical Turk)
- Split huge task into small "Human Intelligence Tasks" (HITs)
- One HIT: Seven sentences, payment of US$ 0.07

- Task: Find relevant text segments (information nuggets)

Hierarchical Summarization Corpus

HIT design



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 the 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:

Hierarchical Summarization Corpus

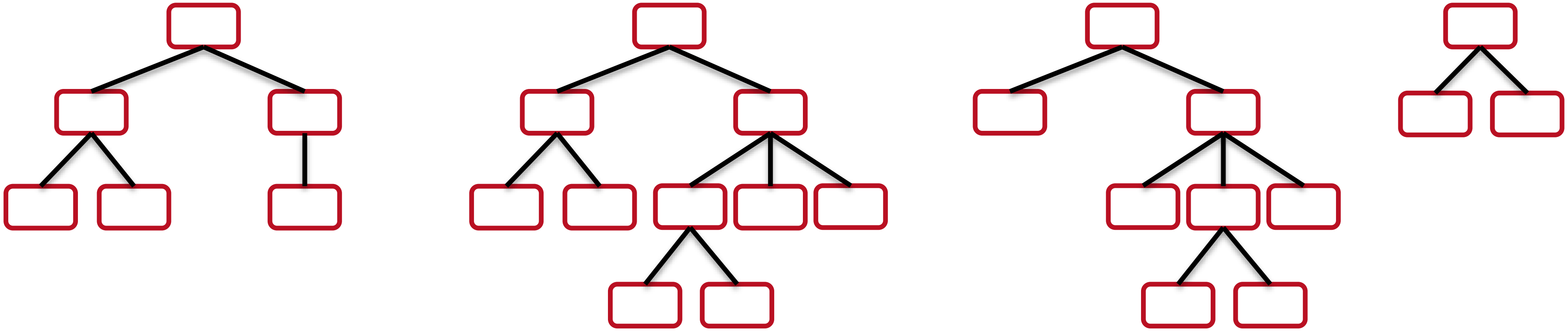
- Each HIT is presented to seven crowd workers
- Workers can annotate nuggets if various length
- Solution: Merge overlapping nuggets

- Only consider nuggets…
  - with at least three tokens
  - that have been annotated by at least three workers

- Result: 4,983 information nuggets (all ten topics)

Hierarchical Summarization Corpus

- We extracted relevant information nuggets for every topic
- How to get to summaries that contain facets?
- Structure the nuggets into hierarchies!



Hierarchical Summarization Corpus

Annotation tool demo

Hierarchical Summarization Corpus

		Facet trees			Average Depth		
Topic	Nuggets	A1	A2	A3	A1	A2	A3
Concerns about religious classes	717	33	81	20	5.42	2.23	3.80
School punishment policy	796	22	29	13	5.45	2.55	6.62
Parents of kids doing drugs	1221	31	139	10	5.35	2.06	7.50
Children’s obesity	445	10	60	11	8.80	2.25	4.45
Sleep problems in preschools	408	17	56	5	7.35	2.25	8.60
Student loans	586	23	44	15	5.92	2.34	4.20
Discipline in elementary school	341	23	48	14	5.13	2.50	3.42
Alternative ADHD treatments	235	14	13	5	3.00	3.77	4.80
Kids with depressions	146	4	33	6	8.50	2.03	6.00

Hierarchical Summarization Corpus

- 10 topics, 3 annotators = 3 hierarchies per topic

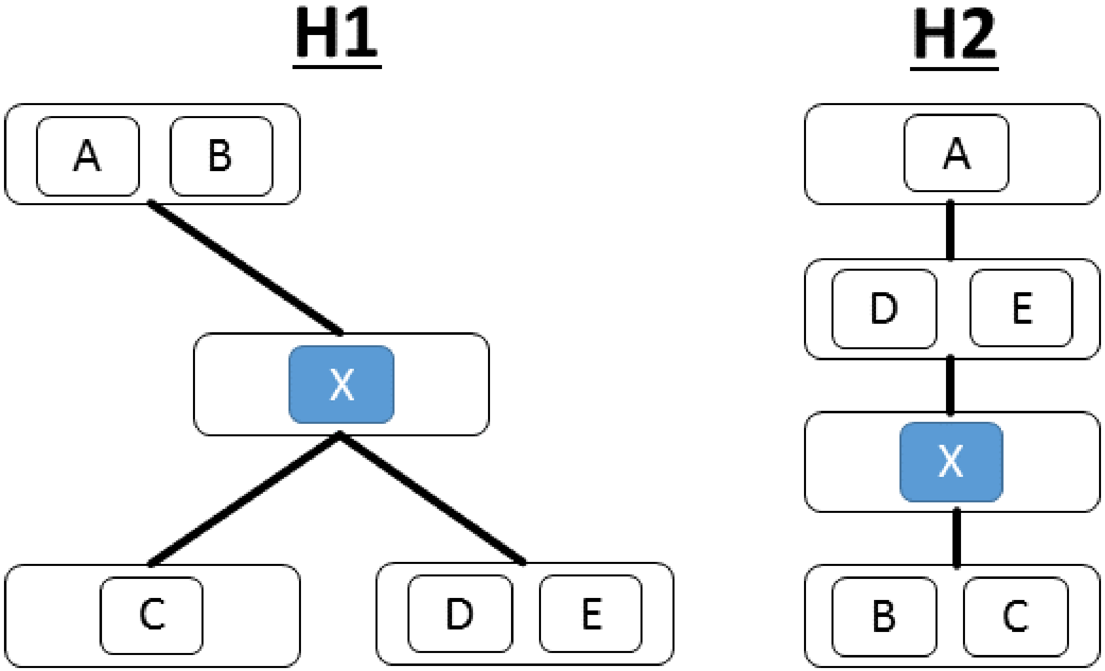
- Is there a best (gold standard) hierarchy?

- First idea: Let annotators discuss their results and create a joined tree
- Problems:
  - Create whole hierarchy together from scratch: Takes a loooong time
  - Start from one result, include ideas of others: Huge bias

- So: Automatic gold standard creation
- Idea: The gold standard is very similar to all three manual hierarchies
- Problem: How to define similarity of hierarchies?

Hierarchical Summarization Corpus

Taxonomy overlap: Compare the sets of super- and sub-components

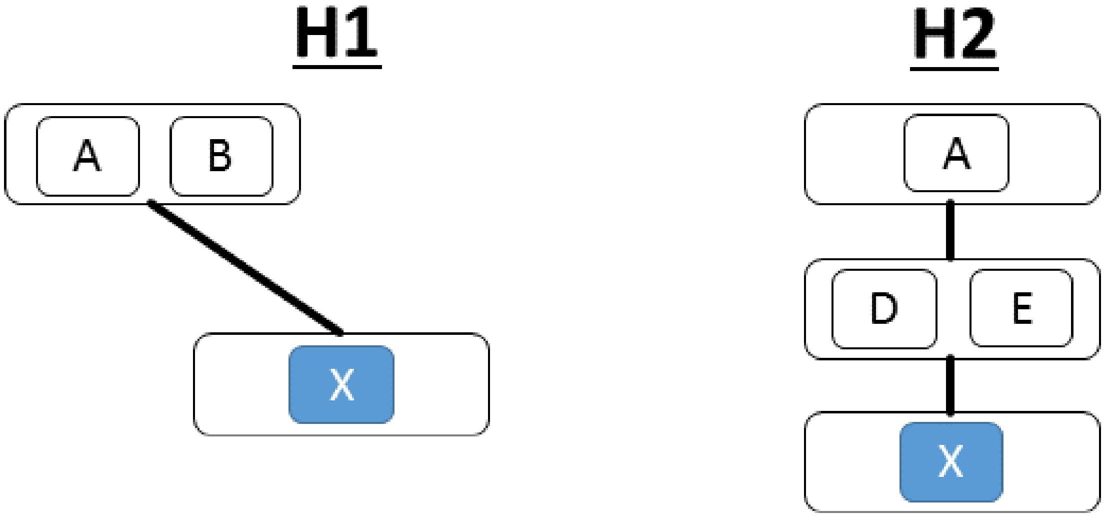


{A, B, C, D, E} versus {A, B, C, D, E}

Perfect match, score 1.0

Hierarchical Summarization Corpus

Hierarchy overlap (HO): Also compare super- and subsets separately

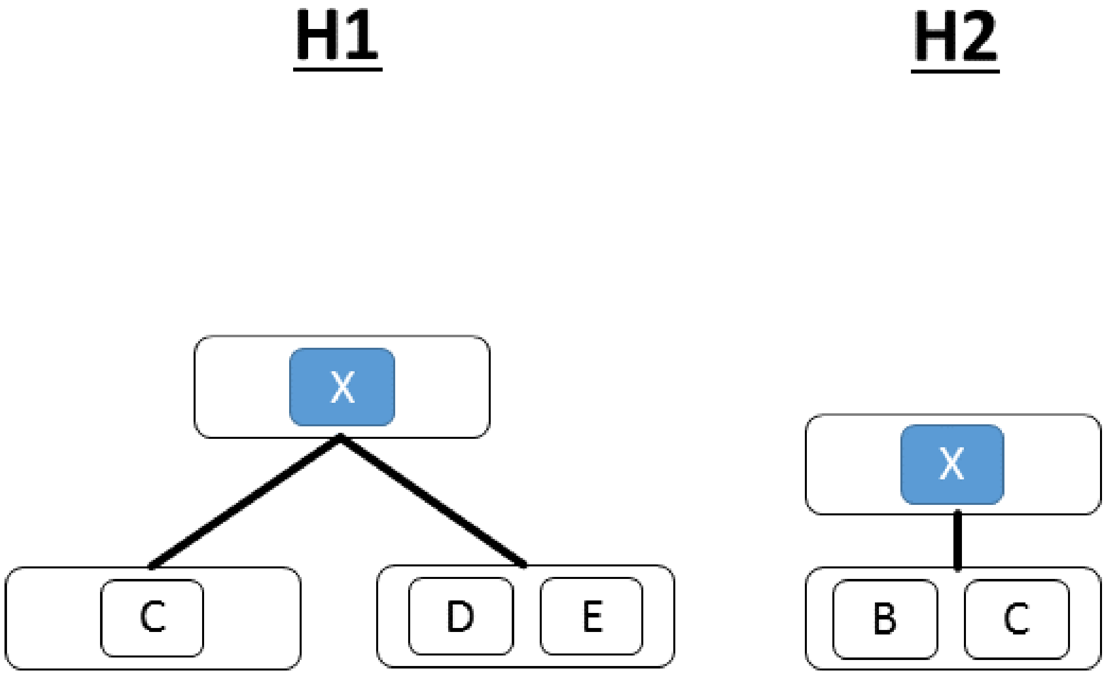


{A, B} versus {A, D, E}

1 joined element {A} of 4 total elements {A, B, D, E} = 0.25

Hierarchical Summarization Corpus

Hierarchy overlap (HO): Also compare super- and subsets separately



{C, D, E} versus {B, C}

1 joined element {C} of 4 total elements {B, C, D, E} = 0.25

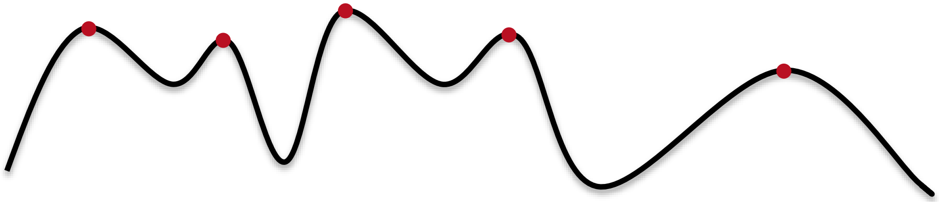
Hierarchical Summarization Corpus

$$HO\left(H_{1,}H_2\right) = a \cdot \underbrace{TO\left(H_{1,}H_2\right)}_{\text{Taxonomy Overlap}} + b \cdot \underbrace{SupO\left(H_{1,}H_2\right)}_{\text{Superset Overlap}} + c \cdot \underbrace{SubO\left(H_{1,}H_2\right)}_{\text{Subset Overlap}}$$

Hierarchical Summarization Corpus

Gold standard algorithm: Local optimization

1. Shuffle list of input nuggets
2. Insert each nugget to maximize HO of whole hierarchy
3. Take out every nugget, insert again at best position
4. Iterate until HO does not increase for any nugget

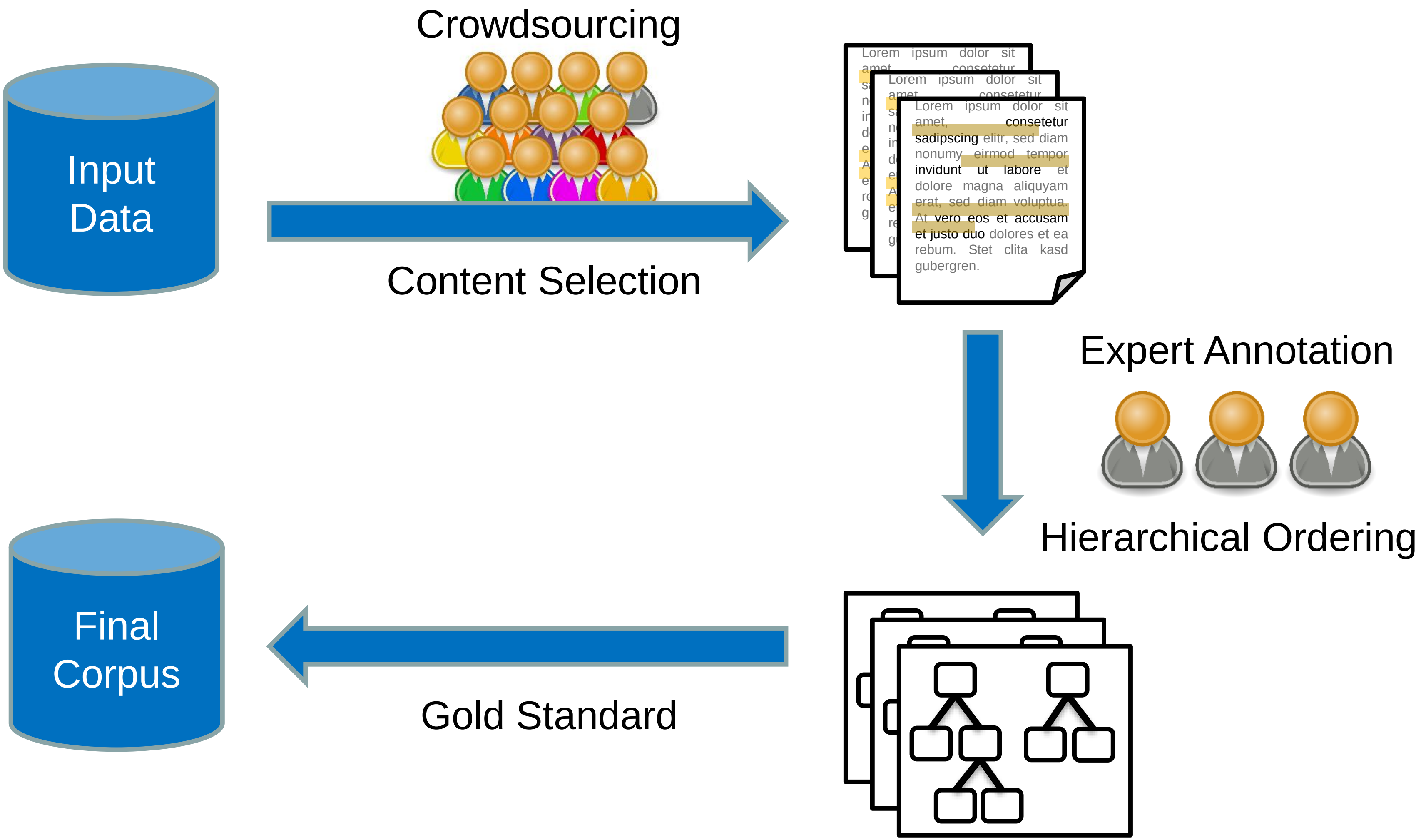


Method finds <u>local</u> optima

Repeat 10 times and take the best result (highest HO)

= gold standard hierarchy

Hierarchical Summarization Corpus



Input Data

Crowdsourcing

Content Selection

Expert Annotation

Hierarchical Ordering

Gold Standard

Final Corpus

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren.

Hierarchical Summarization Corpus

Corpus link: https://github.com/AIPHES/HierarchicalSummarization

Content:

- Source sentences after pre-processing
- HIT template for Amazon Mechanical Turk
- Annotated nugget list
- Curated nugget list as input for hierarchical annotation
- Hierarchy annotation tool, with source code and documentation
- Manual annotated + gold standard hierarchies

References

- https://lemurproject.org/clueweb12/specs.php
- Ivan Habernal et al. "New collection announcement: Focused retrieval over the web." *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016.
- Christopher Tauchmann, Thomas Arnold, Andreas Hanselowski, Christian M. Meyer und Margot Mieskes:
  Beyond Generic Summarization: A Multi-faceted Hierarchical Summarization Corpus of Large Heterogeneous Data. *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC), *erscheint* 2018. Miyazaki, Japan.

Thank you!

- https://lemurproject.org/clueweb12/specs.php
- Ivan Habernal et al. "New collection announcement: Focused retrieval over the web." *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016.
- Christopher Tauchmann, Thomas Arnold, Andreas Hanselowski, Christian M. Meyer und Margot Mieskes:
  Beyond Generic Summarization: A Multi-faceted Hierarchical Summarization Corpus of Large Heterogeneous Data. *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC), *erscheint* 2018. Miyazaki, Japan.