Bringing Structure into Summaries: Crowdsourcing a Benchmark Corpus of Concept Maps

Tobias Falke and **Iryna Gurevych**
Research Training Group AIPHES and UKP Lab
Department of Computer Science, Technische Universität Darmstadt
https://www.aiphes.tu-darmstadt.de

Abstract

Concept maps can be used to concisely represent important information and bring structure into large document collections. Therefore, we study a variant of multi-document summarization that produces summaries in the form of concept maps. However, suitable evaluation datasets for this task are currently missing. To close this gap, we present a newly created corpus of concept maps that summarize heterogeneous collections of web documents on educational topics. It was created using a novel crowdsourcing approach that allows us to efficiently determine important elements in large document collections. We release the corpus along with a baseline system and proposed evaluation protocol to enable further research on this variant of summarization.[1]

1 Introduction

Multi-document summarization (MDS), the transformation of a set of documents into a short text containing their most important aspects, is a long-studied problem in NLP. Generated summaries have been shown to support humans dealing with large document collections in information seeking tasks (McKeown et al., 2005; Maña-López et al., 2004; Roussinov and Chen, 2001). However, when exploring a set of documents manually, humans rarely write a fully-formulated summary for themselves. Instead, user studies (Chin et al., 2009; Kang et al., 2011) show that they note down important keywords and phrases, try to identify relationships between them and organize them accordingly. Therefore, we believe that the study of summarization with similarly structured outputs is an important extension of the traditional task.

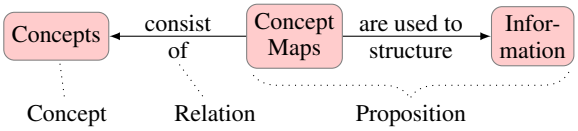

Figure 1: Elements of a concept map.

A representation that is more in line with observed user behavior is a *concept map* (Novak and Gowin, 1984), a labeled graph showing concepts as nodes and relationships between them as edges (Figure 1). Introduced in 1972 as a teaching tool (Novak and Cañas, 2007), concept maps have found many applications in education (Edwards and Fraser, 1983; Roy, 2008), for writing assistance (Villalon, 2012) or to structure information repositories (Briggs et al., 2004; Richardson and Fox, 2005). For summarization, concept maps make it possible to represent a summary concisely and clearly reveal relationships. Moreover, we see a second interesting use case that goes beyond the capabilities of textual summaries: When concepts and relations are linked to corresponding locations in the documents they have been extracted from, the graph can be used to navigate in a document collection, similar to a table of contents. An implementation of this idea has been recently described by Falke and Gurevych (2017).

The corresponding task that we propose is *concept-map-based MDS*, the summarization of a document cluster in the form of a concept map. In order to develop and evaluate methods for the task, gold-standard corpora are necessary, but no suitable corpus is available. The manual creation of such a dataset is very time-consuming, as the annotation includes many subtasks. In particular, an annotator would need to manually identify all concepts in the documents, while only a few of them will eventually end up in the summary.

[1] Available at https://github.com/UKPLab/emnlp2017-cmapsum-corpus

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2951–2961
Copenhagen, Denmark, September 7–11, 2017.

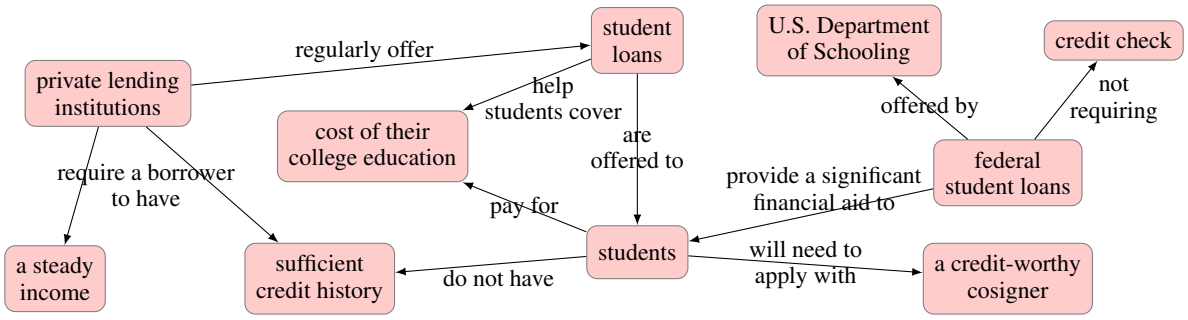

Figure 2: Excerpt from a summary concept map on the topic "*students loans without credit history*".

To overcome these issues, we present a corpus creation method that effectively combines automatic preprocessing, scalable crowdsourcing and high-quality expert annotations. Using it, we can avoid the high effort for single annotators, allowing us to scale to document clusters that are 15 times larger than in traditional summarization corpora. We created a new corpus of 30 topics, each with around 40 source documents on educational topics and a summarizing concept map that is the consensus of many crowdworkers (see Figure 2).

As a crucial step of the corpus creation, we developed a new crowdsourcing scheme called *low-context importance annotation*. In contrast to traditional approaches, it allows us to determine important elements in a document cluster without requiring annotators to read all documents, making it feasible to crowdsource the task and overcome quality issues observed in previous work (Lloret et al., 2013). We show that the approach creates reliable data for our focused summarization scenario and, when tested on traditional summarization corpora, creates annotations that are similar to those obtained by earlier efforts.

To summarize, we make the following contributions: (1) We propose a novel task, concept-map-based MDS (§2), (2) present a new crowdsourcing scheme to create reference summaries (§4), (3) publish a new dataset for the proposed task (§5) and (4) provide an evaluation protocol and baseline (§7). We make these resources publicly available under a permissive license.

2 Task

Concept-map-based MDS is defined as follows: *Given a set of related documents, create a concept map that represents its most important content, satisfies a specified size limit and is connected.*

We define a concept map as a labeled graph showing concepts as nodes and relationships between them as edges. Labels are arbitrary sequences of tokens taken from the documents, making the summarization task extractive. A concept can be an entity, abstract idea, event or activity, designated by its unique label. Good maps should be propositionally coherent, meaning that every relation together with the two connected concepts form a meaningful proposition.

The task is complex, consisting of several interdependent subtasks. One has to extract appropriate labels for concepts and relations and recognize different expressions that refer to the same concept across multiple documents. Further, one has to select the most important concepts and relations for the summary and finally organize them in a graph satisfying the connectedness and size constraints.

3 Related Work

Some attempts have been made to automatically construct concept maps from text, working with either single documents (Zubrinic et al., 2015; Villalon, 2012; Valerio and Leake, 2006; Kowata et al., 2010) or document clusters (Qasim et al., 2013; Zouaq and Nkambou, 2009; Rajaraman and Tan, 2002). These approaches extract concept and relation labels from syntactic structures and connect them to build a concept map. However, common task definitions and comparable evaluations are missing. In addition, only a few of them, namely Villalon (2012) and Valerio and Leake (2006), define summarization as their goal and try to compress the input to a substantially smaller size. Our newly proposed task and the created large-cluster dataset fill these gaps as they emphasize the summarization aspect of the task.

For the subtask of selecting summary-worthy concepts and relations, techniques developed for traditional summarization (Nenkova and McKeown, 2011) and keyphrase extraction (Hasan and Ng, 2014) are related and applicable. Approaches

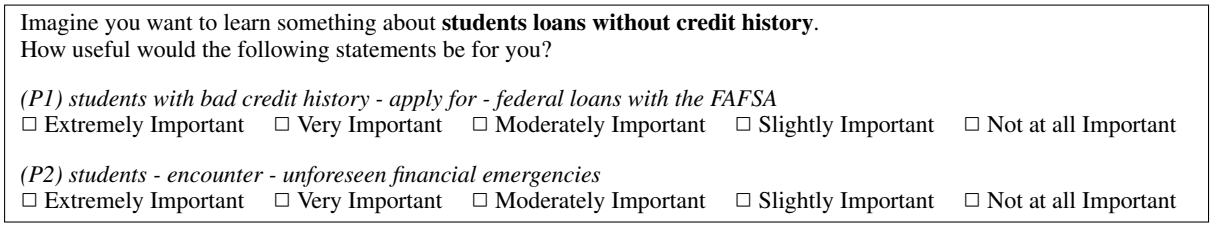

Figure 3: Likert-scale crowdsourcing task with topic description and two example propositions.

that build graphs of propositions to create a summary (Fang et al., 2016; Li et al., 2016; Liu et al., 2015; Li, 2015) seem to be particularly related, however, there is one important difference: While they use graphs as an intermediate representation from which a textual summary is then generated, the goal of the proposed task is to create a graph that is directly interpretable and useful for a user. In contrast, these intermediate graphs, e.g. AMR, are hardly useful for a typical, non-linguist user.

For traditional summarization, the most well-known datasets emerged out of the DUC and TAC competitions.[2] They provide clusters of news articles with gold-standard summaries. Extending these efforts, several more specialized corpora have been created: With regard to size, Nakano et al. (2010) present a corpus of summaries for large-scale collections of web pages. Recently, corpora with more heterogeneous documents have been suggested, e.g. (Zopf et al., 2016) and (Benikova et al., 2016). The corpus we present combines these aspects, as it has large clusters of heterogeneous documents, and provides a necessary benchmark to evaluate the proposed task.

For concept map generation, one corpus with human-created summary concept maps for student essays has been created (Villalon et al., 2010). In contrast to our corpus, it only deals with single documents, requires a two orders of magnitude smaller amount of compression of the input and is not publicly available .

Other types of information representation that also model concepts and their relationships are knowledge bases, such as Freebase (Bollacker et al., 2009), and ontologies. However, they both differ in important aspects: Whereas concept maps follow an open label paradigm and are meant to be interpretable by humans, knowledge bases and ontologies are usually more strictly typed and made to be machine-readable. Moreover, approaches to automatically construct them from text typically try to extract as much information as possible, while we want to summarize a document.

[2] duc.nist.gov, tac.nist.gov

4 Low-Context Importance Annotation

Lloret et al. (2013) describe several experiments to crowdsource reference summaries. Workers are asked to read 10 documents and then select 10 summary sentences from them for a reward of $0.05. They discovered several challenges, including poor work quality and the subjectiveness of the annotation task, indicating that crowdsourcing is not useful for this purpose.

To overcome these issues, we introduce a new task design, *low-context importance annotation*, to determine summary-worthy parts of documents. Compared to Lloret et al.'s approach, it is more in line with crowdsourcing best practices, as the tasks are simple, intuitive and small (Sabou et al., 2014) and workers receive reasonable payment (Fort et al., 2011). Most importantly, it is also much more efficient and scalable, as it does not require workers to read all documents in a cluster.

4.1 Task Design

We break down the task of importance annotation to the level of single propositions. The goal of our crowdsourcing scheme is to obtain a score for each proposition indicating its importance in a document cluster, such that a ranking according to the score would reveal what is most important and should be included in a summary. In contrast to other work, we do not show the documents to the workers at all, but provide only a description of the document cluster's topic along with the propositions. This ensures that tasks are small, simple and can be done quickly (see Figure 3).

In preliminary tests, we found that this design, despite the minimal context, works reasonably on our focused clusters on common educational topics. For instance, consider Figure 3: One can easily say that P1 is more important than P2 without reading the documents.

We distinguish two task variants:

Likert-scale Tasks Instead of enforcing binary importance decisions, we use a 5-point Likert-scale to allow more fine-grained annotations. The obtained labels are translated into scores (5..1) and the average of all scores for a proposition is used as an estimate for its importance. This follows the idea that while single workers might find the task subjective, the consensus of multiple workers, represented in the average score, tends to be less subjective due to the "wisdom of the crowd". We randomly group five propositions into a task.

Comparison Tasks As an alternative, we use a second task design based on pairwise comparisons. Comparisons are known to be easier to make and more consistent (Belz and Kow, 2010), but also more expensive, as the number of pairs grows quadratically with the number of objects.[3] To reduce the cost, we group five propositions into a task and ask workers to order them by importance per drag-and-drop. From the results, we derive pairwise comparisons and use TrueSkill (Herbrich et al., 2007), a powerful Bayesian rank induction model (Zhang et al., 2016), to obtain importance estimates for each proposition.

4.2 Pilot Study

To verify the proposed approach, we conducted a pilot study on Amazon Mechanical Turk using data from TAC2008 (Dang and Owczarzak, 2008). We collected importance estimates for 474 propositions extracted from the first three clusters[4] using both task designs. Each Likert-scale task was assigned to 5 different workers and awarded $0.06. For comparison tasks, we also collected 5 labels each, paid $0.05 and sampled around 7% of all possible pairs. We submitted them in batches of 100 pairs and selected pairs for subsequent batches based on the confidence of the TrueSkill model.

Quality Control Following the observations of Lloret et al. (2013), we established several measures for quality control. First, we restricted our tasks to workers from the US with an approval rate of at least 95%. Second, we identified low quality workers by measuring the correlation of each worker's Likert-scores with the average of the other four scores. The worst workers (at most 5% of all labels) were removed.

In addition, we included trap sentences, similar as in (Lloret et al., 2013), in around 80 of the tasks. In contrast to Lloret et al.'s findings, both an obvious trap sentence (*This sentence is not important*) and a less obvious but unimportant one (*Barack Obama graduated from Harvard Law*) were consistently labeled as unimportant (1.08 and 1.14), indicating that the workers did the task properly.

Agreement and Reliability For Likert-scale tasks, we follow Snow et al. (2008) and calculate agreement as the average Pearson correlation of a worker's Likert-score with the average score of the remaining workers.[5] This measure is less strict than exact label agreement and can account for close labels and high- or low-scoring workers. We observe a correlation of 0.81, indicating substantial agreement. For comparisons, the majority agreement is 0.73. To further examine the reliability of the collected data, we followed the approach of Kiritchenko and Mohammed (2016) and simply repeated the crowdsourcing for one of the three topics. Between the importance estimates calculated from the first and second run, we found a Pearson correlation of 0.82 (Spearman 0.78) for Likert-scale tasks and 0.69 (Spearman 0.66) for comparison tasks. This shows that the approach, despite the subjectiveness of the task, allows us to collect reliable annotations.

Peer Evaluation In addition to the reliability studies, we extrinsically evaluated the annotations in the task of summary evaluation. For each of the 58 peer summaries in TAC2008, we calculated a score as the sum of the importance estimates of the propositions it contains. Table 1 shows how these peer scores, averaged over the three topics, correlate with the manual responsiveness scores assigned during TAC in comparison

Peer Scoring	Pearson	Spearman
Modified Pyramid	0.4587	0.4676
ROUGE-2	0.3062	0.3486
Crowd-Likert	0.4589	0.4196
Crowd-Comparison	0.4564	0.3761

Table 1: Correlation of peer scores with manual responsiveness scores on TAC2008 topics 01-03.

[3] Even with intelligent sampling strategies, such as the active learning in CrowdBT (Chen et al., 2013), the number of pairs is only reduced by a constant factor (Zhang et al., 2016).

[4] D0801A-A, D0802A-A, D0803A-A

[5] As workers are not consistent across all items, we create five meta-workers by sorting the labels per proposition.

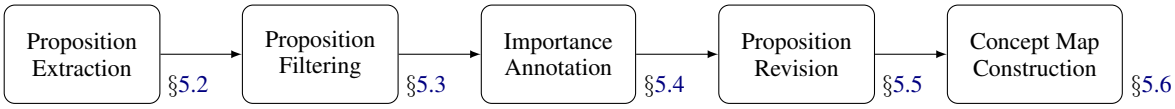

Figure 4: Steps of the corpus creation (with references to the corresponding sections).

to ROUGE-2 and Pyramid scores.[6] The results demonstrate that with both task designs, we obtain importance annotations that are similarly useful for summary evaluation as pyramid annotations or gold-standard summaries (used for ROUGE).

Conclusion Based on the pilot study, we conclude that the proposed crowdsourcing scheme allows us to obtain proper importance annotations for propositions. As workers are not required to read all documents, the annotation is much more efficient and scalable as with traditional methods.

5 Corpus Creation

This section presents the corpus construction process, as outlined in Figure 4, combining automatic preprocessing, scalable crowdsourcing and high-quality expert annotations to be able to scale to the size of our document clusters. For every topic, we spent about $150 on crowdsourcing and 1.5h of expert annotations, while just a single annotator would already need over 8 hours (at 200 words per minute) to read all documents of a topic.

5.1 Source Data

As a starting point, we used the DIP corpus (Habernal et al., 2016), a collection of 49 clusters of 100 web pages on educational topics (e.g. bullying, homeschooling, drugs) with a short description of each topic. It was created from a large web crawl using state-of-the-art information retrieval. We selected 30 of the topics for which we created the necessary concept map annotations.

5.2 Proposition Extraction

As concept maps consist of propositions expressing the relation between concepts (see Figure 1), we need to impose such a structure upon the plain text in the document clusters. This could be done by manually annotating spans representing concepts and relations, however, the size of our clusters makes this a huge effort: 2288 sentences per topic (69k in total) need to be processed. Therefore, we resort to an automatic approach.

[6] Correlations for ROUGE and Pyramid are lower than reported in TAC since we only use 3 topics instead of all 48.

The Open Information Extraction paradigm (Banko et al., 2007) offers a representation very similar to the desired one. For instance, from

Students with bad credit history should not lose hope and apply for federal loans with the FAFSA.

Open IE systems extract tuples of two arguments and a relation phrase representing propositions:

(s. with bad credit history, should not lose, hope)
(s. with bad credit history, apply for, federal loans with the FAFSA)

While the relation phrase is similar to a relation in a concept map, many arguments in these tuples represent useful concepts. We used Open IE 4[7], a state-of-the-art system (Stanovsky and Dagan, 2016) to process all sentences. After removing duplicates, we obtained 4137 tuples per topic.

Since we want to create a gold-standard corpus, we have to ensure that we produce high-quality data. We therefore made use of the confidence assigned to every extracted tuple to filter out low quality ones. To ensure that we do not filter too aggressively (and miss important aspects in the final summary), we manually annotated 500 tuples sampled from all topics for correctness. On the first 250 of them, we tuned the filter threshold to 0.5, which keeps 98.7% of the correct extractions in the unseen second half. After filtering, a topic had on average 2850 propositions (85k in total).

5.3 Proposition Filtering

Despite the similarity of the Open IE paradigm, not every extracted tuple is a suitable proposition for a concept map. To reduce the effort in the subsequent steps, we therefore want to filter out unsuitable ones. A tuple is suitable if it (1) is a correct extraction, (2) is meaningful without any context and (3) has arguments that represent proper concepts. We created a guideline explaining when to label a tuple as suitable for a concept map and performed a small annotation study. Three annotators independently labeled 500 randomly sam-

[7] https://github.com/knowitall/openie

pled tuples. The agreement was 82% ($\kappa = 0.60$). We found tuples to be unsuitable mostly because they had unresolvable pronouns, conflicting with (2), or arguments that were full clauses or propositions, conflicting with (3), while (1) was mostly taken care of by the confidence filtering in §5.2.

Due to the high number of tuples we decided to automate the filtering step. We trained a linear SVM on the majority voted annotations. As features, we used the extraction confidence, length of arguments and relations as well as part-of-speech tags, among others. To ensure that the automatic classification does not remove suitable propositions, we tuned the classifier to avoid false negatives. In particular, we introduced class weights, improving precision on the negative class at the cost of a higher fraction of positive classifications. Additionally, we manually verified a certain number of the most uncertain negative classifications to further improve performance. When 20% of the classifications are manually verified and corrected, we found that our model trained on 350 labeled instances achieves 93% precision on negative classifications on the unseen 150 instances. We found this to be a reasonable trade-off of automation and data quality and applied the model to the full dataset.

The classifier filtered out 43% of the propositions, leaving 1622 per topic. We manually examined the 17k least confident negative classifications and corrected 955 of them. We also corrected positive classifications for certain types of tuples for which we knew the classifier to be imprecise. Finally, each topic was left with an average of 1554 propositions (47k in total).

5.4 Importance Annotation

Given the propositions identified in the previous step, we now applied our crowdsourcing scheme as described in §4 to determine their importance. To cope with the large number of propositions, we combine the two task designs: First, we collect Likert-scores from 5 workers for each proposition, clean the data and calculate average scores. Then, using only the top 100 propositions[8] according to these scores, we crowdsource 10% of all possible pairwise comparisons among them. Using TrueSkill, we obtain a fine-grained ranking of the 100 most important propositions.

[8]We also add all propositions with the same score as the 100th, yielding 112 propositions on average.

For Likert-scores, the average agreement over all topics is 0.80, while the majority agreement for comparisons is 0.78. We repeated the data collection for three randomly selected topics and found the Pearson correlation between both runs to be 0.73 (Spearman 0.73) for Likert-scores and 0.72 (Spearman 0.71) for comparisons. These figures show that the crowdsourcing approach works on this dataset as reliably as on the TAC documents.

In total, we uploaded 53k scoring and 12k comparison tasks to Mechanical Turk, spending $4425.45 including fees. From the fine-grained ranking of the 100 most important propositions, we select the top 50 per topic to construct a summary concept map in the subsequent steps.

5.5 Proposition Revision

Having a manageable number of propositions, an annotator then applied a few straightforward transformations that correct common errors of the Open IE system. First, we break down propositions with conjunctions in either of the arguments into separate propositions per conjunct, which the Open IE system sometimes fails to do. And second, we correct span errors that might occur in the argument or relation phrases, especially when sentences were not properly segmented. As a result, we have a set of high quality propositions for our concept map, consisting of, due to the first transformation, 56.1 propositions per topic on average.

5.6 Concept Map Construction

In this final step, we connect the set of important propositions to form a graph. For instance, given the following two propositions

(student, may borrow, Stafford Loan)
(the student, does not have, a credit history)

one can easily see, although the first arguments differ slightly, that both labels describe the concept *student*, allowing us to build a concept map with the concepts *student*, *Stafford Loan* and *credit history*. The annotation task thus involves deciding which of the available propositions to include in the map, which of their concepts to merge and, when merging, which of the available labels to use. As these decisions highly depend upon each other and require context, we decided to use expert annotators rather than crowdsource the subtasks.

Annotators were given the topic description and the most important, ranked propositions. Using

Corpus	Cluster	Cluster Size	Docs	Doc. Size	Rel. Std.
This work	30	97,880 ± 50,086.2	40.5 ± 6.8	2,412.8 ± 3,764.1	1.56
DUC 2006	50	17,461 ± 6,627.8	25.0 ± 0.0	729.2 ± 542.3	0.74
DUC 2004	50	6,721 ± 3,017.9	10.0 ± 0.0	672.1 ± 506.3	0.75
TAC 2008A	48	5,892 ± 2,832.4	10.0 ± 0.0	589.2 ± 480.3	0.82

Table 2: Topic clusters in comparison to classic corpora (size in token, mean with standard deviation).

a simple annotation tool providing a visualization of the graph, they could connect the propositions step by step. They were instructed to reach a size of 25 concepts, the recommended maximum size for a concept map (Novak and Cañas, 2007). Further, they should prefer more important propositions and ensure connectedness. When connecting two propositions, they were asked to keep the concept label that was appropriate for both propositions. To support the annotators, the tool used ADW (Pilehvar et al., 2013), a state-of-the-art approach for semantic similarity, to suggest possible connections. The annotation was carried out by graduate students with a background in NLP after receiving an introduction into the guidelines and tool and annotating a first example.

If an annotator was not able to connect 25 concepts, she was allowed to create up to three synthetic relations with freely defined labels, making the maps slightly abstractive. On average, the constructed maps have 0.77 synthetic relations, mostly connecting concepts whose relation is too obvious to be explicitly stated in text (e.g. between *Montessori teacher* and *Montessori education*).

To assess the reliability of this annotation step, we had the first three maps created by two annotators. We casted the task of selecting propositions to be included in the map as a binary decision task and observed an agreement of 84% ($\kappa = 0.66$). Second, we modeled the decision which concepts to join as a binary decision on all pairs of common concepts, observing an agreement of 95% ($\kappa = 0.70$). And finally, we compared which concept labels the annotators decided to include in the final map, observing 85% ($\kappa = 0.69$) agreement. Hence, the annotation shows substantial agreement (Landis and Koch, 1977).

6 Corpus Analysis

In this section, we describe our newly created corpus, which, in addition to having summaries in the form of concept maps, differs from traditional summarization corpora in several aspects.

6.1 Document Clusters

Size The corpus consists of document clusters for 30 different topics. Each of them contains around 40 documents with on average 2413 tokens, which leads to an average cluster size of 97,880 token. With these characteristics, the document clusters are 15 times larger than typical DUC clusters of ten documents and five times larger than the 25-document-clusters (Table 2). In addition, the documents are also more variable in terms of length, as the (length-adjusted) standard deviation is twice as high as in the other corpora. With these properties, the corpus represents an interesting challenge towards real-world application scenarios, in which users typically have to deal with much more than ten documents.

Genres Because we used a large web crawl as the source for our corpus, it contains documents from a variety of genres. To further analyze this property, we categorized a sample of 50 documents from the corpus. Among them, we found professionally written articles and blog posts (28%), educational material for parents and kids (26%), personal blog posts (16%), forum discussions and comments (12%), commented link collections (12%) and scientific articles (6%).

Textual Heterogeneity In addition to the variety of genres, the documents also differ in terms of language use. To capture this property, we follow Zopf et al. (2016) and compute, for every topic, the average Jensen-Shannon divergence between the word distribution of one document and the word distribution in the remaining documents. The higher this value is, the more the language differs between documents. We found the average divergence over all topics to be 0.3490, whereas it is 0.3019 in DUC 2004 and 0.3188 in TAC 2008A.

6.2 Concept Maps

As Table 3 shows, each of the 30 reference concept maps has exactly 25 concepts and between 24 and 28 relations. Labels for both concepts and

	per Map	Token	Character
Concepts	25.0 ± 0.0	3.2 ± 0.5	22.0 ± 4.1
Relations	25.2 ± 1.3	3.2 ± 0.5	17.1 ± 2.6

Table 3: Size of concept maps (mean with std).

relations consist on average of 3.2 tokens, whereas the latter are a bit shorter in characters.

To obtain a better picture of what kind of text spans have been used as labels, we automatically tagged them with their part-of-speech and determined their head with a dependency parser. Concept labels tend to be headed by nouns (82%) or verbs (15%), while they also contain adjectives, prepositions and determiners. Relation labels, on the other hand, are almost always headed by a verb (94%) and contain prepositions, nouns and particles in addition. These distributions are very similar to those reported by Villalon et al. (2010) for their (single-document) concept map corpus.

Analyzing the graph structure of the maps, we found that all of them are connected. They have on average 7.2 central concepts with more than one relation, while the remaining ones occur in only one proposition. We found that achieving a higher number of connections would mean compromising importance, i.e. including less important propositions, and decided against it.

7 Baseline Experiments

In this section, we briefly describe a baseline and evaluation scripts that we release, with a detailed documentation, along with the corpus.

Baseline Method We implemented a simple approach inspired by previous work on concept map generation and keyphrase extraction. For a document cluster, it performs the following steps:

1. Extract all NPs as potential concepts.
2. Merge potential concepts whose labels match after stemming into a single concept.
3. For each pair of concepts co-occurring in a sentence, select the tokens in between as a potential relation if they contain a verb.
4. If a pair of concepts has more than one relation, select the one with the shortest label.
5. Assign an importance score to every concept and rank them accordingly.

Metric	Pr	Re	F1
Strict Match	.0006	.0026	.0010
METEOR	.1512	.1949	.1700
ROUGE-2	.0603	.1798	.0891

Table 4: Baseline performance on test set.

6. Find a connected graph of 25 concepts with high scores among all extracted concepts and relations.

For (5), we trained a binary classifier to identify the important concepts in the set of all potential concepts. We used common features for keyphrase extraction, including position, frequency and length, and Weka's Random Forest (Hall et al., 2009) implementation as the model. At inference time, we use the classifiers confidence for a positive classification as the score.

In step (6), we start with the full graph of all extracted concepts and relations and use a heuristic to find a subgraph that is connected, satisfies the size limit of 25 concepts and has many high-scoring concepts: We iteratively remove the weakest concept until only one connected component of 25 concepts or less remains, which is used the summary concept map. This approach guarantees that the concept map is connected, but might not find the subset of concepts that has the highest total importance score.

Evaluation Metrics In order to automatically compare generated concept maps with reference maps, we propose three metrics.[9] As a concept map is fully defined by the set of its propositions, we can compute precision, recall and F1-scores between the two proposition set of generated and reference map. A proposition is represented as the concatenation of concept and relation labels. *Strict Match* compares them after stemming and only counts exact and complete matches. Using *METEOR* (Denkowski and Lavie, 2014), we offer a second metric that takes synonyms and paraphrases into account and also scores partial matches. And finally, we compute *ROUGE-2* (Lin, 2004) between the concatenation of all propositions from the maps. These automatic measures might be complemented with a human evaluation.

Results Table 4 shows the performance of the baseline. An analysis of the single pipeline steps

[9]For precise definitions of the metrics, please refer to the published scripts and accompanying documentation.

revealed major bottlenecks of the method and challenges of the task. First, we observed that around 76% of gold concepts are covered by the extraction (step 1+2), while the top 25 concepts (step 5) only contain 17% of the gold concepts. Hence, content selection is a major challenge, stemming from the large cluster sizes in the corpus. Second, while also 17% of gold concepts are contained in the final maps (step 6), scores for strict proposition matching are low, indicating a poor performance of the relation extraction (step 3). The propagation of these errors along the pipeline contributes to overall low scores.

8 Conclusion

In this work, we presented low-context importance annotation, a novel crowdsourcing scheme that we used to create a new benchmark corpus for concept-map-based MDS. The corpus has large-scale document clusters of heterogeneous web documents, posing a challenging summarization task. Together with the corpus, we provide implementations of a baseline method and evaluation scripts and hope that our efforts facilitate future research on this variant of summarization.

Acknowledgments

We would like to thank Teresa Botschen, Andreas Hanselowski and Markus Zopf for their help with the annotation work and Christian Meyer for his valuable feedback. 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.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 2670–2676, Hyderabad, India.

Anja Belz and Eric Kow. 2010. Comparing Rating Scales and Preference Judgements in Language Evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 7–16, Trim, Ireland.

Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. 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.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2009. Freebase. In *Compilation Proceedings of the International Conference on Management Data & 27th Symposium on Principles of Database Systems*, pages 1247–1250, Vancouver, Canada.

Geoffrey Briggs, David A. Shamma, Alberto J. Cañas, Roger Carff, Jeffrey Scargle, and Joseph D. Novak. 2004. Concept Maps Applied to Mars Exploration Public Outreach. In *Concept Maps: Theory, Methodology, Technology. Proceedings of the First International Conference on Concept Mapping*, pages 109–116, Pamplona, Spain.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise Ranking Aggregation in a Crowdsourced Setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202, Rome, Italy.

George Chin, Olga A. Kuchar, and Katherine E. Wolf. 2009. Exploring the Analytical Processes of Intelligence Analysts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20, Boston, MA, USA.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of the First Text Analysis Conference*, pages 1–16, Gaithersburg, MD, USA.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA.

John Edwards and Kym Fraser. 1983. Concept Maps as Reflectors of Conceptual Understanding. *Research in Science Education*, 13(1):19–26.

Tobias Falke and Iryna Gurevych. 2017. GraphDocExplore: A Framework for the Experimental Comparison of Graph-based Document Exploration Techniques. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Yimai Fang, Haoyue Zhu, Ewa Muszyńska, Alexander Kuhnle, and Simone Teufel. 2016. A Proposition-Based Abstractive Summariser. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 567–578, Osaka, Japan.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37(2):413–420.

Ivan Habernal, Maria Sukhareva, Fiana Raiber, Anna Shtok, Oren Kurland, Hadar Ronen, Judit Bar-Ilan, and Iryna Gurevych. 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.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1262–1273, Baltimore, MD, USA.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill(TM): A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 19*, pages 569–576, Vancouver, Canada.

Youn-Ah Kang, Carsten Görg, and John T. Stasko. 2011. How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):570–583.

Svetlana Kiritchenko and Saif M. Mohammed. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best-Worst Scaling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, CA, USA.

Juliana H. Kowata, Davidson Cury, and Maria Claudia Silva Boeres. 2010. Concept Maps Core Elements Candidates Recognition from Text. In *Concept Maps: Making Learning Meaningful. Proceedings of the 4th International Conference on Concept Mapping*, pages 120–127, Vina del Mar, Chile.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Wei Li. 2015. Abstractive Multi-document Summarization with Semantic Information Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1908–1913, Lisbon, Portugal.

Wei Li, Lei He, and Hai Zhuge. 2016. Abstractive News Summarization based on Event Semantic Link Network. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 236–246, Osaka, Japan.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward Abstractive Summarization Using Semantic Representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado.

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

Manuel J. Maña-López, Manuel de Buenaga, and José M. Gómez-Hidalgo. 2004. Multidocument Summarization: An Added Value to Clustering in Interactive Retrieval. *ACM Transactions on Information Systems*, 22(2):215–241.

Kathleen McKeown, Rebecca J. Passonneau, David K. Elson, Ani Nenkova, and Julia Hirschberg. 2005. Do summaries help? A Task-Based Evaluation of Multi-Document Summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 210–217, Salvador, Brazil.

Masahiro Nakano, Hideyuki Shibuki, Rintaro Miyazaki, Madoka Ishioroshi, Koichi Kaneko, and Tatsunori Mori. 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'10)*, pages 3125–3131, Valletta, Malta.

Ani Nenkova and Kathleen R. McKeown. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2):103–233.

Joseph D. Novak and Alberto J. Cañas. 2007. Theoretical Origins of Concept Maps, How to Construct Them, and Uses in Education. *Reflecting Education*, 3(1):29–42.

Joseph D. Novak and D. Bob Gowin. 1984. *Learning How to Learn*. Cambridge University Press, Cambridge.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1341–1351, Sofia, Bulgaria.

Iqbal Qasim, Jin-Woo Jeong, Jee-Uk Heu, and Dong-Ho Lee. 2013. Concept map construction from text documents using affinity propagation. *Journal of Information Science*, 39(6):719–736.

Kanagasabai Rajaraman and Ah-Hwee Tan. 2002. Knowledge discovery from texts: A Concept Frame Graph Approach. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 669–671, McLean, VA, USA.

Ryan Richardson and Edward A. Fox. 2005. Using concept maps as a cross-language resource discovery tool for large documents in digital libraries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, page 415, Denver, CO, USA.

Dmitri G. Roussinov and Hsinchun Chen. 2001. Information navigation on the web by clustering and summarizing query results. *Information Processing & Management*, 37(6):789–816.

Debopriyo Roy. 2008. Using Concept Maps for Information Conceptualization and Schematization in Technical Reading and Writing Courses: A Case Study for Computer Science Majors in Japan. In *IEEE International Professional Communication Conference (IPCC 2008)*, pages 1–12, Montreal, Canada.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 859–866, Reykjavik, Iceland.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a Large Benchmark for Open Information Extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, TX, USA.

Alejandro Valerio and David B. Leake. 2006. Jump-Starting Concept Map Construction with Knowledge Extracted from Documents. In *Proceedings of the 2nd International Conference on Concept Mapping*, pages 296–303, San José, Costa Rica.

Jorge J. Villalon. 2012. *Automated Generation of Concept Maps to Support Writing*. PhD Thesis, University of Sydney, Australia.

Jorge J. Villalon, Rafael A. Calvo, and Rodrigo Montenegro. 2010. Analysis of a Gold Standard for Concept Map Mining - How Humans Summarize Text Using Concept Maps. In *Proceedings of the 4th International Conference on Concept Mapping*, pages 14–22, Vina del Mar, Chile.

Xiaohang Zhang, Guoliang Li, and Jianhua Feng. 2016. Crowdsourced Top-k Algorithms: An Experimental Evaluation. *Proceedings of the Very Large Databases Endowment*, 9(8):612–623.

Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler. 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.

Amal Zouaq and Roger Nkambou. 2009. Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project. *IEEE Transactions on Knowledge and Data Engineering*, 21(11):1559–1572.

Krunoslav Zubrinic, Ines Obradovic, and Tomo Sjekavica. 2015. Implementation of method for generating concept map from unstructured text in the Croatian language. In *23rd International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 220–223, Split, Croatia.