







Aspect-Driven Structuring of Historical Dutch Newspaper Archives

Hermann Kroll¹(✉) , Christin Katharina Kretz² , Mirjam Cuper³ ,
Bill Matthias Thang¹, and Wolf-Tilo Balke¹ 

¹ TU Braunschweig, Braunschweig, Germany
{kroll,balke}@ifis.cs.tu-bs.de

² TH Köln (University of Applied Sciences), Cologne, Germany
christin.kretz@th-koeln.de

³ KB, National Library of the Netherlands, Hague, The Netherlands
mirjam.cuper@kb.nl

Abstract. Digital libraries oftentimes provide access to historical newspaper archives via keyword-based search. Historical figures and their roles are particularly interesting cognitive access points in historical research. Structuring and clustering news articles would allow more sophisticated access for users to explore such information. However, real-world limitations such as the lack of training data, licensing restrictions and non-English text with OCR errors make the composition of such a system difficult and cost-intensive in practice. In this work we tackle these issues with the showcase of the National Library of the Netherlands by introducing a role-based interface that structures news articles on historical persons. In-depth, component-wise evaluations and interviews with domain experts highlighted our prototype's effectiveness and appropriateness for a real-world digital library collection.

Keywords: Historical News Archives · Exploration · Digital Libraries

1 Introduction

Users of digital libraries featuring historical news articles conduct a variety of information interactions such as task planning or searching for and working with information objects [20]. In historical research, historical figures and especially their roles are particularly interesting cognitive access points [19]. Kumpulainen et al. [19] state the need for supporting historians' research by providing domain-specific tools tailored to their needs. One crucial task of researchers is the creation of sub-corpora to answer their research questions [29]. However, finding these sub-corpora, especially when researchers are unfamiliar with the searched historical persons, can be challenging for two reasons. First, the huge size of news article archives might be overwhelming. Second, posing and finding suitable keyword queries to browse such archives is difficult.

Advances in Natural Language Processing (NLP) lead to historic news systems with novel access paths for users to engage with their content [12]. A variety

of such digital library projects has been proposed in the past, e.g., NewsEye [17], ANNO [27], impresso [11], or Cuper’s work [5] (see Sect. 2 for a detailed discussion). However, those systems usually rely either on manual curation [5] or at least domain-specific training examples for every implemented step [11]. In contrast, our work bypasses manual curation and the collection of domain-specific training data by utilizing data from Wikipedia (structure information with text examples). This paper introduces a novel system that automatically structures historical news articles on persons and provides an aspect-driven interface to explore a library’s content. The central idea is that a person has different roles (e.g., *writer*, *politician*, *military person*) and each role has different aspects (e.g., *early life*, *political career*, *actions*). Our system should, at best, automatically create sub-corpora for each role and aspect to support research on historical persons. However, traditional methods introduced in the NLP domain typically rely on hand-crafted training data and sometimes artificial benchmarks [30]. We tackled the challenges faced by an actual digital library, namely the National Library of the Netherlands, Koninklijke Bibliotheek (KB) (<https://www.kb.nl>). Here, no hand-crafted training data and benchmarks were available. Moreover, the library imposed several real-world constraints: (1) The data was available in Dutch, whereas NLP methods are often available in English only. (2) The news articles were based on OCR-scanned newspapers, and hence, came with typical OCR issues (such as incorrect letters or broken paragraphs). (3) The data came with a license prohibiting sending data to APIs like ChatGPT [1].

In addition to those constraints, which are typically not the target in NLP research, we observed an understudied [20] corpus of non-English but Dutch news articles. Our overall goal was thus to build a real-world system that overcomes the typical constraints of a typical digital library. In this work, we therefore strive to support users’ data-driven process planning by structuring news articles concerning historical figures by their respective roles. Our prototypical system operates on real data of the KB and bases on automatically generated training data from Wikipedia. We expected our system to help users in the formulation of research questions on the provided data of historical persons.

To tackle our overall research question *How can a digital library design effective access paths to explore their collection?*, we made the following contributions: (1) We discuss and demonstrate how we overcome a digital library’s real-world restrictions and constraints (see Sect. 3). (2) We present an effective method for automatically structuring news articles by employing structural background information from Wikipedia with the use case of news articles on historical figures. (3) We evaluate our prototypical system step-by-step and via interviews with five domain experts. Code is available at GitHub¹ and Software Heritage².

¹ <https://github.com/HermannKroll/AspectDrivenNewsStructuring>.

² <https://archive.softwareheritage.org/swh:1:dir:13457c154ed7ad1f571e353c1edf2f87db61b0ae>.

2 Related Work

Related work for our research objective falls into the following categories: (1) related digital library news archive retrieval systems, (2) processing Dutch texts via language models, and (3) text summarization methods.

Digital Library Systems on News Articles. Structuring and exploring news has been a topic of wide research, e.g., summarization [30], the evolution of terms [25], fake news detection [35], clustering [24] and many more. For instance, [24] clusters news articles based on their similarity to pre-computed categories using SVMs. Kumpulainen et al. [19] identified roles of historical persons, relationships between them, and in general, named entities as important cognitive access points to historical documents. Clustering similar news articles has been explored in several concrete applications with real digital library constraints, e.g., NewsEye [17] or ANNO [27]. Another example is the Swiss-Luxembourgish project impresso [11] which utilizes NLP methods like named entity recognition, word embeddings, n-gram search, and information extraction to provide additional information on historical news articles. The KB has developed the Delpher platform: News articles were digitized by OCR tools and Delpher provides a user interface to navigate through their historical newspaper collections. Beyond the traditional keyword-based search, they aimed to organize a part of the KB’s newspaper collection differently from the standard search interface [5]. Additionally, the KB manually created subject pages that give more background information on certain topics and related newspapers³. Our work’s goal was to structure the KB’s news articles automatically, at least as much as possible, while meeting the KB’s real-world constraints.

Dutch Language Models. Many language models were trained and evaluated on English corpora. Exceptions were models trained in a *multilingual setting* [9,23,37] or ones having been *trained for Dutch*: BERTje [36] is a Dutch BERT [9] model which outperforms the multilingual version [23] of BART [21]. RobBERT [6] is a Dutch RoBERTa model which outperformed BERTje on the sentiment analysis task as well as both BERTje and mBERT on the relative pronoun prediction tasks. A newer version of the model (RobBERT-2022) [7] with a newer Dutch training corpus also outperformed BERTje and RobBERT on the sentiment analysis task. We used the RobBERT-2022 for our text classification.

Text Summarization. The task of text summarization is to produce a concise natural language summary. Nowadays, general-application sequence-to-sequence language models can be fine-tuned to solve the text summarization task, e.g., UniLM [10], T5 [31], BART [21], PEGASUS [40]. Another option is using large language models (LLMs) [41] and prompting in this context. Models like [40] or [21] are restricted to 512 tokens or less, meaning their input must be shorter than 512 tokens. So-called longformer models surpass this restriction by allowing up to 16k tokens as their input, e.g., [2,30,39]. Beyond some remarkable examples like Estonian [16] and Romanian news summarization [28], text summarization models are trained in English [2,21,30,31,40], but they can be fine-tuned for

³ <https://www.delpher.nl/thema/geschiedenis/tweede-wereldoorlog>.



Fig. 1. User interface of our system, URL: <https://narrative.pubpharm.de/news>.

other languages (here Dutch). The goal of this work was to summarize several articles in a single summary, so the *multi-document summarization* task. PRIMERA [30] is an LED-based [2] state-of-the-art model (ACL2022) for this task. It outperformed single-document models [2, 21, 40] in different scenarios (news and scientific documents). That is why we used PRIMERA in this work.

3 Conception and Data Acquisition

Our overall goal was to structure news articles to support corresponding research questions on single persons. Each new article consists of a title, the textual content, the release date, and the publishing newspaper. From our viewpoint, each person might have different roles $r \in \mathcal{R}$ (e.g., politician, writer, artist) that come with different aspects r_A (e.g., political career, novels, awards).

Discussion of the Library's Constraints. In brief, we faced the following constraints: (1) The texts stem from OCR-scanned news articles using ABBYY Finereader, (2) texts were written in Dutch (no translation was available), (3) prohibition against sending data to third parties, (4) forced linking to the Delpher system and restriction to show only snippets of the actual data (160 characters at max), and (5) no curated training data for any of our sub-tasks. Those requirements forced us to exclude automated translation services like DeepL and AI assistants like ChatGPT by design. Especially the lack of training data prevented the usage of straightforward approaches like training text classification models. We would have had to collect data for roles and aspects, manually label news articles, and then train classification models. However, creating such data would be cost intensive.

That is why we headed for a different approach: We used the Dutch Wikipedia to gather texts describing different persons, their roles, and the roles' aspects. First, Wikipedia organizes text into different sections describing different *aspects* of entries. Second, Wikipedia enriches an item's text through so-called info boxes that provide structured information, e.g., whether it is a person and has some

Table 1. Statistics for our news article collection: #Art. describes how many articles before filtering and #FArt. after filtering were retrieved for each person.

Person Name (Title/Synonyms)	Role	Life	#Art.	#FArt.
Winston Churchill (Sir Churchill)	politician	1874-1965	47k	8463
Leopold III van België (Leo. III, prins Leo.)	king	1901-1983	26k	1677
Wilhelmina van Oranje-Nassau (prinses Wilhelmina, koningin Wilhelmina)	queen	1880-1962	257k	9416
Jannetje Schaft (Hannie Schaft)	resistance	1920-1945	2056	34
Dwight Eisenhower (majoor-generaal E., Generaal E., president E.)	politician	1890-1969	114k	21k
Anne Frank	war victim	1929-1945	11k	1
Frans Goedhart (Pieter 't Hoen)	resistance	1904-1990	4105	560
Simon Vestdijk	writer	1898-1971	5544	1453
Franklin Roosevelt (president Roosevelt)	politician	1882-1945	165k	16k

roles. For our approach, we used the info boxes to determine a person’s role and the Wikipedia texts to learn how certain aspects are described. This approach bypassed the creation of training data, while, however, could cause new problems: It had to be tested if classifiers trained on descriptive Wikipedia texts are transferable and generalizable to Dutch news.

Prototype (User Interface). In constructing the system interface, we strive to cater to McCay-Peet et al.’s [26] five facets supporting serendipity in digital environments: interfaces filled with various information (*trigger-rich*), showcase relationships between information objects (*enables connections*), visual cues (*highlights triggers*), *enables exploration* and provides unanticipated or surprising information (*leads to the unexpected*). Our method’s goal was to (1) derive the roles of a person (trigger-rich and exploration) and (2) classify whether a news article’s content belongs to one of the role’s aspects (connections and unexpected). We used multi-document summarization for each aspect to help users quickly access what is written in the corresponding articles. Users should be able to select different persons and one of the person’s known roles. Then users could navigate through different aspects of that role, see a summary for each aspect and a list of articles classified as belonging to that role’s aspect (see Fig. 1 for the systems’ screenshot and URL). A click on an article forwards users to Delpher.

Historical News Data from the KB. We used a subset of the KB’s data for building our system since the KB collected news articles from the 17th century to the recent past. We selected articles for nine famous persons in relation to the second world war with various roles because the KB’s Delpher has identified the second world war as a topic users were interested in. We harvested relevant articles by querying for the name and title/pseudonym (see Table 1 for statistics). We only selected items from the newspaper collection with the type ‘article’. Then we only kept articles where $\geq 90\%$ of the text was found in a Dutch dictionary, as recommended in [34], to remove noisy and low-quality OCR-scanned

data. We also excluded newspapers published by fascist organizations or German authorities with a national socialist agenda. Articles for each person should, on the one hand, carry enough information about the person and, on the other hand, stem from the time when the person was alive. That is why we applied the following additional filters: (1) A news article’s release date must be in the corresponding person’s life span, (2) the article content must be longer than 100 words, and (3) a person’s partial name (e.g., *Frank* for *Anne Frank*) should be mentioned at least three times. Especially the time constraint did filter nearly all articles, except one, of Anne Frank because they were published after her death.

4 System Implementation

Wikipedia Processing. As already mentioned, we used the Wikipedia info boxes to derive a person’s role. The information was linked to Wikipedia categories which were organized in a taxonomy, e.g., *British politician* is a specialization of a *politician*. In our context, we understood a person’s occupation as a role. We crawled the Dutch *occupation* categories and derived a list of occupations (in sum 30k distinct ones). Then, we iterated through the Dutch Wikipedia XML dumps (March 2023), parsed the info boxes, checked whether a property of the info box was linked to one of those occupations, and if so, we extracted the corresponding page’s summary (introduction) and sections plus all occupations. In sum, we derived 259k person pages. While reviewing the pages, we observed many very short pages, e.g., including a brief summary or a single section. However, our goal was to find frequent aspects of well-described roles. So, we removed all pages that (1) had a less than 150 characters summary, or (2) had < 3 sections. Note that we disregarded sections with less than 100 characters and sections that only contained references/literature by using a hand-crafted list. This filtering reduced the number of person pages to 61k. With that, we obtained roles plus thousands of Wikipedia pages for each. Wikipedia sections should, at best, describe one unit of information belonging to a certain aspect of a person.

However, Wikipedia was crafted collaboratively, i.e., through human editing meaning section titles are usually not-canonicalized. For instance, *life*, *background*, and *curriculum vitae/resume* describe the same, or at least a very similar, aspect of a person. To face this concern, we designed a canonicalization step to cluster semantically similar sections. We applied a pre-trained sentence transformer model (BERT-base-dutch-cased) using the S-BERT Library [32], capable of embedding semantically similar sentences closely in its vector space. To embed a section, we embedded all of its sentences and then computed the mean vector over all sentence vectors. Next, we averaged all section vectors with equal titles (e.g., background sections). Finally, we compared those sections vectors pairwise using the cosine similarity. If two vectors’ similarity exceeds a certain pre-defined threshold, we consider those section titles as semantically equivalent. We then computed the transitive closure to determine the set of semantically equivalent section titles, i.e., if a-b and b-c are merged, we also merged a-c. To retain a high precision, we used a similarity threshold of 0.95 in our system.

Aspect Mining and Classification. Next, we mined frequent role aspects by counting how often the aspect (section title or any other section title from that same cluster) was used across all persons of a role (e.g., *writers*). We then computed a relative support, e.g., 0.2 means that 20% of all *writers* have aspect (section title or any similar title) *background*. We defined a minimum absolute (to ensure enough text examples for aspect training) and a relative support threshold (to ensure frequency within a role). Given a certain person’s role, we trained a classifier to predict whether a text belongs to one of the role’s aspects. That means we headed for a multi-class classification scenario, e.g., a classifier for role r_1 with aspects a_1, a_2, a_3 must predict one of the aspects, or the negative class (not belonging to the role). First, we retrieved Wikipedia section texts for each aspect. We ensured that each aspect must have at least a minimum number of texts to be considered for training (see aspect mining support threshold). However, some aspects might have more examples than others, which is why we sampled all text examples randomly down to the number of the least frequent aspect, e.g., aspect a_1 and a_2 are sampled down to 100 texts if the least frequent aspect a_3 only has 100 examples. We randomly sampled negative examples (not belonging to the role) from other persons and aspects that do not have the given role r_1 . We sampled as many negative examples as we had positive ones, e.g., 100-100-100 positive (three aspects, 100 texts each) and 300 negative examples.

We fine-tuned the Dutch model RobBERT-2022 [7] for the actual text classification. We split our data into train, validation, and test sets (80-10-10). We performed training on train (5 epochs), and searched for hyperparameters (learning rate [1e-3, 1e-4, 1e-5] and decay [0.1, 0.2]) on validation. We picked the best model concerning validation and macro precision because our classification should prefer precision of all classes over recall. We trained a classifier for each role (occupation category of Wikipedia) that had (1) at least three frequent aspects and (2) belongs to the first two category levels in Wikipedia (to select more general roles like *writer* instead of *British writer*). Note that we removed the category suffixes *naar nationaliteit* and *naar beroep*.

News Article Processing. The next step was applying those classifiers to a historical person’s actual Dutch news articles. However, a news article might include several different topics, thus, classifying the whole text to one role’s aspects could be problematic. So, we computed snippets of the articles that include the person’s name: We split the article’s content into sentences by using NLTK’s [3] sentence split method. We then checked whether a partial name of the person (e.g., *Churchill* or *Winston* for *Winston Churchill*) was included. If so, we considered the sentence relevant and took it and the sentence before and after as additional context information to generate a snippet. The average sentence length computed over Dutch newspaper and Wikipedia articles is 90.3 characters, 3.4% of these sentences have ≤ 19 characters [15]. We only use snippets of three sentences with at least 50 characters to filter out broken or incomplete sentences, corresponding to a minimum average sentence length of ~ 16.7 characters each.

For our selected persons, we identified their roles through the info boxes of the corresponding Wikipedia entry. If a role (e.g., *British minister*) was assigned,

we also considered its super categories (e.g., *minister* and *person*). We always assigned the role *person* to ensure that our approach also worked in cases when a person did not have an info box (in cases of *Wilhelmina* and *Janeetje Schaft*). Having the roles, we applied the corresponding role classifier to every news article snippet of the person. Note that each snippet could be classified as belonging to several aspects of different roles. This was intended because some aspects of different roles might overlap, e.g., a *politician's* and *writer's family* or *early life*.

Our last goal was to summarize those snippets into one summary for the users so that they could quickly grasp how the aspect was described in the news articles. However, to the best of our knowledge, multi-document summarization models were unavailable for Dutch. That is why we decided to apply one of the latest English models, namely PRIMERA [30]. We used a fine-tuned news summary PRIMERA model from HuggingFace. However, to apply PRIMERA, we had to translate the Dutch news article snippets into English with OPUS-MT [33], one of the latest open available translation models. The choice of OPUS-MT over using, e.g., the DeepL API, was again made due to our legal constraints. Afterward, PRIMERA's English summaries were translated back to Dutch with OPUS-MT. We translated Dutch texts sentence-wise to English and vice versa. To generate the summaries, we introduced a parameter k to select how many articles snippets should be summarized. In addition to the summaries, we wanted to display fragments of the article snippets in the user interface, to give our users an idea about the article. For the fragment generation, we identified the position where the person's name was mentioned and displayed the surrounding characters and cut if we exceeded the 160 characters we were allowed to show.

5 Evaluation

We evaluate our system's components individually (clustering, classification, translation, and summarization) and then report our user study's findings.

Clustering. We exported 221 distinct section titles that occurred in at least 100 Wikipedia articles to ensure enough examples for the clustering and classification. We asked three persons to cluster them manually, i.e., whether two titles semantically belong together. When comparing and discussing their clusters, we observed the following patterns: There was a wide range in clustering regarding the granularity. One annotator clustered everything belonging to one's life as one cluster, whereas a second person created clusters for different periods in life such as *youth* with *early life*, *youth and training* and *later life* with *death* and *last years*. The annotators had difficulties distinguishing between titles describing a person, e.g., *author*, and titles describing a person's work, e.g., *novel*. But all annotators differentiated between a *politician* and their *political career*. All three annotators agreed to cluster section titles describing different types of *awards*. The annotators disagreed on whether to cluster military and political careers. War-related titles such as *interbellum* and *after the war* also were regarded with uncertainty regarding them being in separate or the same cluster. In general, the annotators found that some section titles were very hard to cluster as the

Table 2. Evaluation results for our Wikipedia text classifiers. We averaged the number of trained aspects and used training samples. Evaluation metrics are macro averaged.

Setting	#Aspects	#Samples	Precision	Recall	F1	Accuracy
Top-5	7.6 ± 3.83	9999 ± 9520	0.95 ± 0.01	0.94 ± 0.03	0.94 ± 0.02	0.95 ± 0.02
Top-10	6.7 ± 3.16	10246 ± 14078	0.94 ± 0.02	0.93 ± 0.02	0.93 ± 0.02	0.94 ± 0.02
Worst-5	6.4 ± 1.74	1285 ± 230	0.80 ± 0.02	0.79 ± 0.03	0.78 ± 0.03	0.82 ± 0.02
Worst-10	5.7 ± 1.95	1545 ± 1340	0.81 ± 0.02	0.81 ± 0.04	0.80 ± 0.03	0.83 ± 0.02
All (43)	6.35 ± 2.88	4254 ± 8215	0.87 ± 0.05	0.88 ± 0.05	0.87 ± 0.05	0.89 ± 0.04

Table 3. Evaluation results of our article snippet classification. For each person, the number of used snippets, different roles, snippets classified as belonging to one aspect, aspects, and classified snippets per aspect are reported.

Name	#Sni.	#Roles	#Classified	#Aspects	Snippets/Aspects		
					Mean±STD	Min	Max
W. Churchill	48k	15	47k	92	508 ± 1587	1	12172
Leopold III	3192	6	1691	42	40 ± 61	1	332
Wilhelmina	40k	1	231	5	46 ± 24	16	76
Jannetje Schaft	76	1	1	1	1 ± 0	1	1
D. Eisenhower	100k	3	36k	20	1780 ± 6631	1	30568
Anne Frank	9	4	1	1	1 ± 0	1	1
Frans Goedhart	2995	2	1132	12	94 ± 283	1	1031
Simon Vestdijk	4989	3	3368	20	168 ± 462	1	2154
F. Roosevelt	80k	7	40k	50	799 ± 3116	1	21926

titles were ambiguous: *Work* could be associated with a person’s job, but also with its outcome, e.g., paintings of a painter.

In a subsequent discussion, the three annotators also reviewed the system-generated clusters (41 in total) and commented on them. The annotators were content with most of the clustering but found some clusters which they considered too broad (e.g., *work* together with *bibliography*) or included labels which were seemingly unrelated (e.g., *influence* and *scientist*). They remarked on some titles which were not clustered together: *Work* was not in the same cluster as *works*, *military career* and *political career* belong to different clusters, *life* and *young years* were clustered together but *death* and *early years* were in two different clusters. In brief, the clustering quality was acceptable to continue.

Aspect Classification. We evaluated the aspect classification in three ways: (1) Wikipedia classifier quality measured on test sets, (2) article classification statistics, and (3) rated classified snippets in a manual evaluation. For the aspect mining, we selected an absolute support of 100 examples per aspect to ensure enough examples for the subsequent classification, and a relative support of 0.05 to ensure relevance to the role. With that, we trained classifiers for 43 roles that had at least three different frequent aspects. We applied the classifiers to

the Wikipedia test sets to measure the classification quality. The results are reported in Table 2 with additional statistics (avg. training data size, number of aspects). To look at the best and worst performing classifiers (ranked by macro precision to ensure reliable, precise classes), we evaluated five settings: Top-5 classifiers, Top-10 classifiers, Worst-5, Worst-10, and All classifiers. In brief, we concluded two thoughts: The more training data a classifier got, the better its performance was. Top-5 achieved a macro precision of 0.95, while Worst-5 still maintained a precision of 0.8, which we still consider acceptable. The recall was between 0.94 (Top-5) and 0.79 (Worst-10). The number of trained samples was between 10k and 1.2k. However, the deviation was high, e.g., a deviation of 14k for 10k samples. A close look at histograms revealed some outliers, like the role *person* with more than 150k samples. Overall, the classification quality was good.

Table 3 reports statistics on the actual classified news article snippets. For instance, Winston Churchill had up to 15 different roles yielding 47k classified snippets with 92 different role aspects in total. While some role aspects had up to 12k classified snippets, others had only one. Briefly, the number of classified snippets strongly differed between our test persons. Persons like Wilhelmina did not have a role concerning Wikipedia and were hence classified only as a *person*. In our user interface, we show the best-classified snippets plus their summary. That is why we ranked classified snippets by their classification probability and selected the top-5 per person, role, and aspect. From this list with 557 snippets, we randomly sampled 100 entries (role, aspect, snippet) for a manual evaluation.

Three persons rated each entry’s correctness and gave explanations if they tagged an entry as incorrect. Counting the majority votes, we obtained 62 correct and 38 incorrect entries with an inter-rater agreement of 0.33 (Krippendorff’s α [18]) and 0.32 (Fleiss’ κ [13]). Discussing the reasons for the negative ratings revealed that, in many cases, the aspect applied was correctly classified, but the role did not fit. For instance, some aspects like *early life* were way too general to be specific for one role, and hence, deciding whether an *early life* snippet belonged to the role *politician*, *member of the Parliament*, or *writer* was impossible. Annotators were uncertain about how to rate a statement about a person rather than an action performed by the person. Another encountered issue was distinguishing between pairs of roles which could belong together: *journalist* – *writer*, *minister* – *official*, *writer* – *artist*, or *historian* – *writer*. Such a decision strongly influenced the rating of the aspect classification and oftentimes made raters disagree. Some snippets alone were not enough to rate an entry, e.g., if the award *Karlspreis* is given to *writers*.

Translation. We randomly sampled 100 snippets from all news articles. Two native Dutch speakers read the Dutch snippet and the corresponding translated English version. They rated the syntax of the translation (whether it reads well and is syntactically correct) and the factual correctness (whether the translated facts are still correct). For the syntax, the annotators’ ratings for *good-moderate-bad* were 54-28-47 and 38-47-15. The inter-rater agreement was 0.62 (Krippendorff’s α) and 0.39 (Fleiss’ κ). However, annotators often disagreed in rating a snippet as good or moderate. Counting good and moderate together as

Table 4. Summarization evaluation results. The averaged readability scores, the averaged number of summary sentences, and the averaged reading time are shown per k .

Summary@ k	#Sent.	Flesch EN	Flesch NL	Reading Time	Dale-Chall
5	7.7 ± 2.8	70.4 ± 8.8	57.3 ± 9.7	11.9 ± 3.8	9 ± 0.9
10	10.4 ± 4.1	70.1 ± 10.1	57.1 ± 10.2	15.5 ± 4.8	8.8 ± 0.9
20	13.7 ± 5.8	69.8 ± 9.0	56.4 ± 9.0	19.4 ± 5.6	8.8 ± 0.7
30	15.0 ± 5.9	70.2 ± 8.4	56.6 ± 9.4	21.0 ± 6.4	8.7 ± 0.8
40	15.6 ± 6.4	70.5 ± 8.0	57.3 ± 8.8	21.7 ± 6.8	8.7 ± 0.7
50	16.1 ± 6.4	71.3 ± 8.0	57.9 ± 8.3	22.0 ± 6.9	8.7 ± 0.7

one class, we obtained an inter-rater agreement of 0.75 (Krippendorff’s α and Fleiss’ κ), indicating a fair agreement. In brief, between 82–85% of the translation snippets had a moderate or good syntax. Concerning factual correctness, for *correct-incorrect* the ratings were 82–18 and 85–15. The inter-rater agreement was high, 0.85 (Krippendorff’s α and Fleiss’ κ). Discussions with both raters revealed that in most cases, when a snippet was marked as factually incorrect, it was due to a minor error. The translation worked well with older Dutch, apart from some mistakes (such as ‘*Duitschland*’, which was erroneously translated as ‘*Germanland*’). The translation also handled minor OCR errors or spaces.

Summarization. We evaluated the summarization through (1) automated readability scores and (2) a manual evaluation. We only summarized aspects of roles that had at least five classified snippets per person, otherwise, we did not have enough information to show. In addition, we summarized the most probable 20 article snippets based on classification probability as the multi-document summarization model could only process 4096 tokens as its input and will truncate otherwise. With that, we generated 208 summaries in total. Table 4 reports the following averaged measures: The Flesch readability index [14] quantifies reading ease based on word and sentence length and is language-specific. Scores between 50 and 60 indicate fairly difficult text, while scores between 70 and 80 indicate fairly easy text. The reading time indicates the seconds required to read a text, each character taking 14.69 ms [8]. The new Dale-Chall [4] score gives the reading level of a text as a grade indicating the familiarity of persons from that grade with a list of words. Scores from 8.0 to 8.9 correspond to an 11th/12th-grade student’s reading level. We also tested a different number of selected snippets to summarize k , however, except for the number of generated sentences and the required reading time, the scores did not deviate much. We randomly sampled ten summaries plus the 20 snippets used to generate them for three raters. Readability on a *good-average-bad* scale was rated as 0-10-0, 0-10-0 and 1-4-5. The generation quality was hence acceptable. From their discussion of the results we found the following: First, if some snippets supported

parts of the summary, they were nearly cited verbatim. Some snippets of different articles were (nearly) identical. Temporal information in the summaries on dates was often bad because the dates were wrong or messed up. Some summaries included hard context breaks between sentences. Moreover, we observed major issues with factual correctness due to hallucinations. Phrases like “*The New York times reports, click here for more*” were generated but not included in the articles. Further, the model also introduced additional, and often wrong, facts about persons, e.g., dates, events, and actions. We assume that such facts and phrases were already learned in PRIMERA’s pre-training and fine-tuning for news. However, hallucinated facts in summaries were a major issue. A comparable setting (trained on news, tested on other data) found 51–55% factual consistency [42].

User Interviews. We conducted five independent 30-minute interviews with employees of the KB. The interview partners consented to take part in the study voluntarily and have their voices recorded. They were made aware that they could stop and drop out of the interview at any time without consequences. The process (a mail to the investigator) for later deleting user-specific data was also explained. A week before our semi-structured voice-recorded one-on-one interviews in Dutch took place, participants received an email with a video (<https://www.youtube.com/watch?v=0Gzlydjts2E>) explaining our prototypical system and its URL. Each interview tackled the same guide questions concerning general thoughts, encountered problems, (un)clear elements, helpfulness of the system and components (aspects and summaries), and suggested changes.

Results and Findings. The full questions plus answers are available in our GitHub repository. In general, the interviewees were enthusiastic about the interface. They found it well-arranged and clear. The website immediately provided a lot of information and context about the person in question. Some other remarks were that the interface worked intuitively and that clustering articles per subject was a plus point. The option to be directly referred to the complete articles on Delpher was also mentioned positively. The interviewees believed that a website like this could definitely help certain users (such as researchers), mainly because they immediately get some context about a person instead of only a list of articles. However, they all agreed that some human input was still needed to refine the system’s output. The interviewees also provided feedback on the various aspects of the system. They found the roles interesting and a good way to immediately provide information about the person. However, they all had some difficulty in understanding how the roles were chosen as they noticed a lot of overlap between different roles. This led to the question of why these have not been merged (such as the roles *politician* and *politician by party*). Opinions were divided on the aspects. Some found the distinction useful, while others wondered why not all articles belonging to one role were grouped together. They agreed that the multiple labels (clustered section titles of Wikipedia) shown above every aspect should be condensed for clarity for two reasons: The number of labels is unbalanced between aspects, and the sections with many labels caused some confusion, e.g., the aspect with the labels *background*, *biography*, etc. appeared

under every role. The interviewees expected it to only belong to the role *person*. Summaries really posed major issues and worried the interviewees. All unanimously agreed that summaries containing incorrect facts are highly problematic for a library. Some also wondered whether the summary added value.

6 Conclusion

In this work, we demonstrated how a digital library can implement an aspect-driven access path to its news collection. We used Wikipedia to bypass the curation of domain-specific and cost-intensive training data. Moreover, our evaluation verified the method’s effectiveness on real-world data and the system’s value in practice. However, there is still room for improvements, e.g., finding suitable labels for a section cluster, showing and summarizing diverse snippets, and highlighting connections between people. For instance, we could better cater to the requirements of the KB by battling hallucinations in summaries by either fact-checking each sentence against the input summaries and removing unsupported ones or by using an extractive summarization approach [22, 38].

References

1. OpenAI’s ChatGPT. <https://openai.com/blog/chatgpt>
2. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. CoRR abs/2004.05150 (2020). <https://arxiv.org/abs/2004.05150>
3. Bird, S.: NLTK: the natural language toolkit. In: ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006. The Association for Computer Linguistics (2006). <https://doi.org/10.3115/1225403.1225421>
4. Chall, J., Dale, E.: Readability Revisited: The New Dale-Chall Readability Formula. Brookline Books (1995)
5. Cuper, M.: Researching pandemics through time: a Covid-19 inspired data-driven approach to explore historical newspapers. In: Berget, G., Hall, M.M., Brenn, D., Kumpulainen, S. (eds.) TPDFL 2021. LNCS, vol. 12866, pp. 227–231. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86324-1_26
6. Delobelle, P., Winters, T., Berendt, B.: Robbert: a dutch roberta-based language model. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Findings of ACL, vol. EMNLP 2020, pp. 3255–3265 (2020). <https://doi.org/10.18653/v1/2020.findings-emnlp.292>
7. Delobelle, P., Winters, T., Berendt, B.: Robbert-2022: Updating a dutch language model to account for evolving language use. CoRR abs/2211.08192 (2022). <https://doi.org/10.48550/arXiv.2211.08192>
8. Demberg, V., Keller, F.: Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition **109**(2), 193–210 (2008). <https://doi.org/10.1016/j.cognition.2008.07.008>
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT 2019, pp. 4171–4186 (2019). <https://doi.org/10.18653/v1/n19-1423>

10. Dong, L., et al.: Unified language model pre-training for natural language understanding and generation. In: NeurIPS 2019. pp. 13042–13054 (2019). <https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html>
11. Düring, M., Kalyakin, R., Bunout, E., Guido, D.: Impresso inspect and compare. visual comparison of semantically enriched historical newspaper articles. *Inf.* **12**(9), 348 (2021). <https://doi.org/10.3390/info12090348>
12. Ehrmann, M., Bunout, E., Düring, M.: Historical newspaper user interfaces: a review. In: 85th IFLA General Conference and Assembly, Athens, Greece, 24–30 August 2019, pp. 1–24 (2019)
13. Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382 (1971). <https://doi.org/10.1037/h0031619>
14. Flesch, R.F.: A new readability yardstick. *J. Appl. Psychol.* **32**(3), 221–33 (1948)
15. Goldhahn, D., Eckart, T., Quasthoff, U.: Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pp. 759–765. European Language Resources Association (ELRA), May 2012. http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf
16. Härm, H., Alumäe, T.: Abstractive summarization of broadcast news stories for estonian. *Balt. J. Mod. Comput.* **10**(3) (2022). <https://doi.org/10.22364/bjmc.2022.10.3.23>
17. Jean-Caurant, A., Doucet, A.: Accessing and investigating large collections of historical newspapers with the newseye platform. In: JCDL ’20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 531–532. ACM (2020). <https://doi.org/10.1145/3383583.3398627>
18. Krippendorff, K.: Content analysis (1989)
19. Kumpulainen, S., Keskustalo, H., Zhang, B., Stefanidis, K.: Historical reasoning in authentic research tasks: mapping cognitive and document spaces. *J. Assoc. Inf. Sci. Technol.* **71**(2), 230–241 (2020). <https://doi.org/10.1002/asi.24216>
20. Late, E., Kumpulainen, S.: Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources. *J. Documentation* **78**(7), 106–124 (2022). <https://doi.org/10.1108/JD-04-2021-0078>
21. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 7871–7880 (2020). <https://doi.org/10.18653/v1/2020.acl-main.703>
22. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019. pp. 3728–3738. Association for Computational Linguistics (2019). <https://doi.org/10.18653/v1/D19-1387>
23. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics* **8**, 726–742 (2020). https://doi.org/10.1162/tacl_a_00343
24. Maria, N., Silva, M.J.: Building a digital library of web news. In: ECDL 2000, vol. 1923, pp. 344–347 (2000). https://doi.org/10.1007/3-540-45268-0_36
25. Marjanen, J., Pivovarova, L., Zosa, E., Kurunmäki, J.: Clustering ideological terms in historical newspaper data with diachronic word embeddings. In: 5th International Workshop on Computational History, HistoInformatics@TPDL 2019. CEUR

- Workshop Proceedings, vol. 2461, pp. 21–29. CEUR-WS.org (2019). https://ceur-ws.org/Vol-2461/paper_4.pdf
26. McCay-Peet, L., Toms, E.G., Kelloway, E.K.: Development and assessment of the content validity of a scale to measure how well a digital environment facilitates serendipity. *Inf. Res.* **19**(3) (2014). <http://www.informationr.net/ir/19-3/paper630.html>
 27. Müller, C.: A N N O - AUSTRIAN NEWSPAPERS ONLINE: Historische österreichische Zeitungen und Zeitschriften online. Eine Digitalisierungsinitiative der Österreichischen Nationalbibliothek (<http://anno.onb.ac.at/>). K. G. Saur (2004). <https://doi.org/10.1515/9783110944198-023>
 28. Niculescu, M.A., Ruseti, S., Dascalu, M.: Rosummary: control tokens for Romanian news summarization. *Algorithms* **15**(12), 472 (2022). <https://doi.org/10.3390/a15120472>
 29. Pfanztel, E., Oberbichler, S., Marjanen, J., Langlais, P., Hechl, S.: Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. *J. Data Min. Digit. Humanit.* 2021 (2021). <https://doi.org/10.46298/jdmdh.6121>
 30. Phang, J., Zhao, Y., Liu, P.J.: Investigating efficiently extending transformers for long input summarization. *CoRR abs/2208.04347* (2022). <https://doi.org/10.48550/arXiv.2208.04347>
 31. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 140:1–140:67 (2020). <http://jmlr.org/papers/v21/20-074.html>
 32. Reimers, N., Gurevych, I.: Sentence-bert: sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, November 2019. <https://arxiv.org/abs/1908.10084>
 33. Tiedemann, J., Thottingal, S.: OPUS-MT - building open translation services for the World. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, pp. 479–480. European Association for Machine Translation (2020). <https://aclanthology.org/2020.eamt-1.61/>
 34. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of ocr quality on downstream nlp tasks. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH, pp. 484–496. INSTICC, SciTePress (2020). <https://doi.org/10.5220/0009169004840496>
 35. Vogel, I., Jiang, P.: Fake news detection with the new German dataset “German-FakeNC”. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) TPDFL 2019. LNCS, vol. 11799, pp. 288–295. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30760-8_25
 36. de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M.: Bertje: a dutch BERT model. *CoRR abs/1912.09582* (2019). <http://arxiv.org/abs/1912.09582>
 37. Xue, L., et al.: mt5: a massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, pp. 483–498 (2021). <https://doi.org/10.18653/v1/2021.naacl-main.41>
 38. Yadav, A., Ranvijay, R., Yadav, R., Maurya, A.K.: State-of-the-art approach to extractive text summarization: a comprehensive review. *Multimed. Tools Appl.*, 1–63, February 2023. <https://doi.org/10.1007/s11042-023-14613-9>

39. Zaheer, M., et al.: Big bird: transformers for longer sequences. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (2020). <https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html>
40. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In: ICML 2020. 119, pp. 11328–11339, 2020. <http://proceedings.mlr.press/v119/zhang20ae.html>
41. Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K.R., Hashimoto, T.B.: Benchmarking large language models for news summarization. CoRR abs/2301.13848 (2023). <https://doi.org/10.48550/arXiv.2301.13848>
42. Zhang, Z., Elfardy, H., Dreyer, M., Small, K., Ji, H., Bansal, M.: Enhancing multi-document summarization with cross-document graph-based information extraction. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1696–1707. Association for Computational Linguistics, May 2023. <https://aclanthology.org/2023.eacl-main.124>