Collection-Document Summaries

Nils Witt[1](✉), Michael Granitzer[2], and Christin Seifert[2]

[1] ZBW-Leibniz Information Centre for Economics,
Düsternbrooker Weg 120, 24105 Kiel, Germany
`n.witt@zbw.eu`
[2] University of Passau, Innstraße 32, 94032 Passau, Germany
{`Michael.Granitzer,Christin.Seifert`}`@uni-passau.de`
`http://www.zbw.eu`

Abstract. Learning something new from a text requires the reader to build on existing knowledge and add new material at the same time. Therefore, we propose collection-document (CDS) summaries that highlight commonalities and differences between a collection (or a single document) and a single document. We devise evaluation metrics that do not require human judgement, and three algorithms for extracting CDS that are based on single-document keyword-extraction methods. Our evaluation shows that different algorithms have different strengths, e.g. TF-IDF based approach best describes document overlap while the adaption of Rake provides keywords with a broad topical coverage. The proposed criteria and procedure can be used to evaluate document-collection summaries without annotated corpora or provide additional insight in an evaluation with human-generated ground truth.

Keywords: Collection-document summaries · Text summarization

1 Introduction

Learning from educational or scientific texts requires readers to integrate new concepts into their existing background knowledge [1]. In the case of digital libraries this means that every search result has to be judged on existing, new and additional information compared to already acquired knowledge of the user. In digital libraries, this judgment is usually based on explicit summary information about the search result in questions, such as title and abstract and does not include explicit information on what is new and what has already been covered by previous searches or the user's private library. Similarities between a document collection and a document can be measured with a qualitative values (e.g. [4]) and quantitatively judged using single-instance summaries (e.g. [6]). Both, however, cannot provide comprehensive, explicit summaries about what content is covered in both, the collection and the document (commonalities) and what content is new in the document compared to the collection (novelties). In this paper we propose *collection-document summaries*, i.e., textual summaries that stress differences and commonalities between a collection of documents and candidate documents. Concretely, the contributions of this paper are the following:

G. Pasi et al. (Eds.): ECIR 2018, LNCS 10772, pp. 638–643, 2018.
https://doi.org/10.1007/978-3-319-76941-7_56

- We identify requirements for keyword-based collection-document summaries.
- Based on the requirements, we propose evaluation metrics for collection-document summaries that do not require human-centric ground-truth.
- Provide baseline algorithms for collection-document summaries by adapting single-document summarizations methods.

The collection-document summaries are intended to be directly consumed by users, for instance, to help them judge the suitability of a search result. Due to the lack of available training data and the required effort to collect it, we aim for a automatic evaluation that does not require human-centered ground-truth. The focus for collection-document summaries is on transparency for users, but they could also be used as features in recommendation and retrieval algorithms.

2 Related Work

Automatic *text summarization* aims to generate short-length text covering the most important concepts and topics of the text [2]. Text summaries can either be sentences, phrases or keyphrases, and the content of the summary can either be chosen from the document itself (extractive summaries) or generated anew based on the document (abstractive summaries) [5]. Most methods for text summarization either focus on single-documents or adapt single-document methods to multiple documents. *Multi-document* summarization aims to summarize a collection of textual documents [9]. Methods for multi-document summarizaiton include using single-document methods on super-documents (concatenation of all documents from a collection) or averaging the results for single-document methods over the collection [9]. This work relates to multi-document summarization as follows: we also extract summaries for collections of documents, but output the differences and commonalities of a candidate document (not in the collection) to the collection in terms of keyphrases. *Keyphrase extraction* attempts to extract phrases that concessively and most appropriately cover the concepts of the text [3]. In this work, we extend keyphrase extractions to collection-document summaries, by postprocessing the results of two well known-keyphrase extraction methods, namely TextRank [6] and Rake [8] and comparing the results with a simple baseline considering TF-IDF term weights in the vectorspace-model.

3 Collection-Document Summaries

We define Collection-Document Summaries (CDS) as summarization of a collection of documents and a document, representing how the document's content differs and which content it has in common with the collection. Similarly, we can also compare two documents (i.e., as a collection containing a single document). Consider the scenario of a person accessing a new field by reading literature. The reader has already read n papers ($D = \{d_0, ..., d_n\}$) and wants to decide whether to read the paper d_c next. In that scenario the reader is interested to find documents that have some known content to start with and also have some content

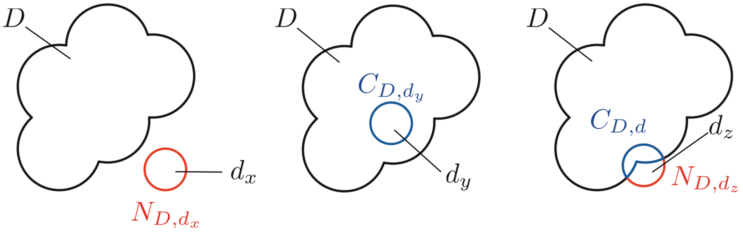

Fig. 1. Types relationships between collections D and documents. Left: d_x differs from D ($C_{D,d_x} = \emptyset$), center: d_y is similar to D ($N_{D,d_y} = \emptyset$), right: collection and document share some concepts, i.e. $N_{D,d_z} \neq \emptyset$ and $C_{D,d_z} \neq \emptyset$

that is new to the reader. In other words, the reader is looking for documents with both, commonalities and novelties:

- **Commonalities:** d_c contains concepts that are also contained in D. These are concepts the reader is already familiar with.
- **Novelties:** d_c contained concepts that are not contained in D. These are the concepts the reader is going to encounter when reading d_c.

Few commonalities and many novelties indicate a big conceptual gap between D and d_c. The reader may have difficulties to grasp the content of d_c. Few novelties and many commonalities on the other hand indicate d_c lacks worthwhile content. We assume the reader is interested in documents with a balanced amount of novelties and commonalities, which may not always be true (e.g. when known concepts are to be revived). While generally, commonalities and novelties as conceptual views on the documents can be represented in multiple ways (e.g. subparts of an ontology), in the remainder of this paper we assume that commonalities and novelties are represented as words. Therefore, we define CDS as follows: *The collection-document summary of a collection D and a document d (i.e. $\Delta(D, d)$) is the pair $(C_{D,d}, N_{D,d})$ where $C_{D,d}$ represents the common keywords and $N_{D,d}$ the novel keywords of documentd with respect to D.*

Figure 1 shows three types of relations between collections and documents. We will motivate and discuss desired properties of CDS and then propose according evaluation measures for these properties in the next section.

- **Comparability:** a document d_c similar to the collection D should introduce no (or only few) new keywords, i.e., if the content of d is already covered by the collection this should be reflected in the keywords.
- **Differentiability:** a document d_c that is not similar to the collection D introduces new keywords, i.e., the difference should be visible by viewing the keywords.
- **Diversity:** the keywords of either commonalities or novelties of a document should cover all concepts that the document deals with.
- **Specificity:** the keywords of either commonalities or novelties should be specific rather than abstract, e.g. *university education* is preferred over *education*.

- **Utility:** The above criteria are necessary but not sufficient, as they do not assess whether the results are meaningful for users. Generally, it requires humans to assess whether CDS are meaningful for a given task, standard metrics to measure utility are precision, recall and F1 w.r.t. to the human-annotated ground-truth.

4 Experiments

In the experiments we evaluated three different algorithms for DCS with the criteria presented in Sect. 3. Source code and data sets are publicly available[1].

Data Sets. The data set consists of 140,341 scientific papers from the economic domain available in the digital collection EconStor[2]. The data set contains information about author, paper abstract, paper type, publication year, venue and a set of JEL-classification codes[3] in meta-data fields. For our experiments we selected those papers that have an abstract and at least one JEL code assigned resulting in 67,813 documents. We annotated phrase candidates of at most 3 terms using the phrase collocation detection described by Mikolov et al. [7]. We constructed artificial user collections D containing k documents with the following property: All documents in the collection must have at least n JEL codes in common, where $n \in \{1, \ldots, 8\}$ is an agreement parameter. Additionally, we randomly generate documents d_x and d_y with the following properties: d_x must have all JEL codes present in the collection and d_y must not have any of the collection's JEL codes (cf. Fig. 1 for a visualization). We chose the agreement on JEL codes for constructing the collection and determining the similarities because JEL codes provide an abstract, topical view on the documents, comprise multiple topics and are high-quality human-annotated meta-data fields. The parameters were set to $k = 10$ and $n = 5$ in our experiments.

Algorithms. For the simple baseline, ΔTF, we rank the words of a documents by their TF-IDF score and select the upper 20% of that list. For ΔTR, we applied TextRank [6] on the documents, keeping the top 20% of the words. We used the TextRank implementation of the Python summa library. For $\Delta Rake$, we used Rake [8] from the Python library rake_nltk. All the algorithms create a set of keywords for a single document. The keywords for a collection were derived using the set union operator for all documents in a collection. The *commonalities* $C(D, d)$ were calculated as the set intersection between the keywords of the collection D and the document d. *Novelties* $N(D, d)$ were calculated by subtracting the set of keywords of the collection D from the set of keywords from the document d (Table 1).

Evaluation Measures. We measure **Comparability** and **Differentiability** as the size of the keyword overlap: $\frac{kw_m(d_c) \cap kw_m(D)}{kw_m(d_c)}$, where, in the case of comparability $d_c = d_x$ and in the case of differentiability $d_c = d_y$ (cf. Fig. 1). $kw_m(d)$

---

[1] https://doi.org/10.5281/zenodo.1133311.

[2] https://www.econstor.eu, last accessed 2017-10-27.

[3] https://www.aeaweb.org/jel/guide/jel.php, last accessed 2017-10-27.

Table 1. Example keywords.

$\Delta Rake$	ΔTF	ΔTR
unanticipated reform, major change, cultural conditions, mothers income, order births, favorable institutional, strong labor market attachment	compensate, hampered, births, essentially, unanticipated, unfavorable, mothers	fully compensated, essential incentives, mothers, largely driven, earlier

Table 2. Overview of results. Showing mean and variance aggregated for all measures

Method	Keywords per doc	Comparability	Differentiability	Specificity	Diversity
$\Delta Rake$	13.5 ± 6.6	0.37 ± 0.04	**0.10** ± 0.03	2.9% ± 0.4%	**0.60** ± 0.10
ΔTF	6.2 ± 2.9	**0.50** ± 0.06	0.13 ± 0.04	**1.2%** ± 0.3%	0.15 ± 0.03
ΔTR	3.6 ± 2.2	0.45 ± 0.03	0.17 ± 0.09	2.2% ± 0.8%	0.18 ± 0.06
Samples		100	100	500	10,000

is the number of keywords extracted by method m on the document d. For **Diversity** we construct a binary JEL code-keyword matrix (M) for each keyword extraction algorithm on the entire data set. Each entry m_{ij} in M indicates whether a specific JEL code i occurs in at least t documents for which keyword i has also been extracted. The parameter t is set to 10 for Rake, 20 for TF-IDF and 5 for TextRank in the experiments. These values were obtained by manual optimization. Thus, the columns of M contain representations of keywords in terms of JEL codes. In a second step, given a candidate document, keywords are extracted and their respective columns of M are combined by logical OR yielding a vector v. The ground-truth JEL codes for the document are compared to the candidate vector v using Jaccard similarity. To measure the **Specificity** we generate two disjoint collections, i.e. two collections that do not share any JEL code. Afterwards the keywords of both collections are extracted and the set intersection and set symmetric difference are computed. The intuition behind this is, that, since the two collections share no JEL codes, they are topically different. Hence, for keyword extractors that generate specific keywords the intersection should be empty. Keywords in the intersection are expected to be unspecific. This measure is normalize by the amount of generated keywords. We divide the intersection size by the size of the symmetric difference.

5 Results

The results of our experiments are summarized in Table 2. We see that the average amount of keywords produced varies considerably, with Rake producing too many keywords given the assumption that the results should be consumed by people. ΔTF scores best at *Comparability* and achieves proper results in *Differentiability*, leading to the larges gap between these two related measures.

That means that ΔTF is the preferred method to model the assumption depicted in Fig. 1. Presumably, $\Delta Rake$'s bad *Comparability* performance can be partially explained by its much larger number of unique keywords, which makes matching keywords less probable whereby the comparability score drops. $\Delta Rake$'s bad *Specificity* performance is surprising, as it has the largest repertoire of keywords available, which should allow it extract specific keywords. ΔTF on the other hand performs much better albeit its much smaller keyword repertoire. High *Diversity* scores indicate that the keywords a method extracts are good classification features to predict the JEL codes of documents. This is the measure where $\Delta Rake$ excels, due to its multi-token keywords (cf. Table 2) and probably also because of the higher keyword per document count.

6 Summary

We have introduced the notion of collection-document summaries and identified criteria by which the quality of those summaries can be measured. Furthermore, we have conducted experiments with three keyword extraction methods. The applied keyword extraction methods are state-of-the-art methods for single document summarization, and therefore should be considered a lower bound baseline for collection-document summarization. Future work includes the devision of new algorithms, for instance by combining $\Delta Rake$ (best diversity) and ΔTF (best comparability) and an evaluation of the methods on a human-generated ground-truth to answer the question about the utility of the extracted keywords.

References

1. Eddy, M.D.: Fallible or inerrant? a belated review of the constructivist's bible. Br. J. Hist. Sci. **37**(1), 93–98 (2004). Jan Golinski, making natural knowledge: Constructivism and the history of science. Cambridge history of science
2. Gambhir, M., Gupta, V.: Recent automatic text summarization techniques: a survey. Artif. Intell. Rev. **47**(1), 1–66 (2017)
3. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: a survey of the state of the art. In: ACL, vol. 1, pp. 1262–1273 (2014)
4. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the New Zealand Computer Science Research Student Conference, pp. 49–56 (2008)
5. Mani, I.: Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)
6. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain (2004)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the International Conference on Neural Information Processing Systems, NIPS 2013, vol. 2, pp. 3111–3119 (2013)
8. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic Keyword Extraction from Individual Documents. Wiley, New York (2010). pp. 1–20
9. Verma, R.M., Lee, D.: Extractive summarization: limits, compression, generalized model and heuristics. CoRR abs/1704.05550 (2017)