WikiAsp: A Dataset for Multi-domain Aspect-based Summarization

Hiroaki Hayashi[1] **Prashant Budania**[2] **Peng Wang**[2]
Chris Ackerson[2] **Raj Neervannan**[2] **Graham Neubig**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]AlphaSense
{hiroakih,gneubig}@cs.cmu.edu
{pbudania,pwang,cackerson,rneervannan}@alpha-sense.com

Abstract

Aspect-based summarization is the task of generating focused summaries based on specific points of interest. Such summaries aid efficient analysis of text, such as quickly understanding reviews or opinions from different angles. However, due to large differences in the type of aspects for different domains (e.g., sentiment, product features), the development of previous models has tended to be domain-specific. In this paper, we propose WikiAsp,[1] a large-scale dataset for multi-domain aspect-based summarization that attempts to spur research in the direction of open-domain aspect-based summarization. Specifically, we build the dataset using Wikipedia articles from 20 different domains, using the section titles and boundaries of each article as a proxy for aspect annotation. We propose several straightforward baseline models for this task and conduct experiments on the dataset. Results highlight key challenges that existing summarization models face in this setting, such as proper pronoun handling of quoted sources and consistent explanation of time-sensitive events.

1 Introduction

Aspect-based summarization is a subtask of summarization that aims to provide targeted summaries of a document from different perspectives (Titov and McDonald, 2008; Lu et al., 2009; Wang and Ling, 2016; Yang et al., 2018; Angelidis and Lapata, 2018). Unlike generic summarization, this gives more concise summaries that are separated according to specific points of interest, allowing readers to fulfill focused information needs more easily and quickly. However, existing aspect-based summarization work is somewhat narrowly focused; for example, a great majority of the work focuses specifically on the domain of product or restaurant reviews. In contrast, generic summarization models are tested on a much wider variety of genres, from newswire (Nallapati et al., 2016; Grusky et al., 2018), to academic papers (Kang et al., 2018; Kedzie et al., 2018), to movie scripts (Gorinski and Lapata, 2015). For each genre, the types and characteristics of aspects that will need to be touched upon in a good summary will differ greatly.

One natural source of such multi-domain articles is Wikipedia, and the section boundaries and titles in each article form natural annotations of aspects and corresponding text. There have recently been a number of attempts to generate the *lead* section of Wikipedia articles from the linked external sites in the reference section (Liu et al., 2018; Fan et al., 2019; Liu and Lapata, 2019a), an approach that does not explicitly consider the different aspects covered by the article. Perez-Beltrachini (2019) also examine domain differences in Wikipedia text summarization. However, existing datasets and analyses lack structure, broad domain coverage, or both. We argue that (1) generating *structured* summaries is of inherent interest, as these will allow humans consuming the information to browse specific aspects of interest more readily, and (2) the structure will *vary across domains*, with different domains demonstrating very different characteristics.

In this paper, we construct a dataset for multi-domain aspect-based summarization that allows us to train models for this unique variety of summarization task, and examine the challenges posed therein. Figure 1 illustrates the overview of our task. Specifically, we turn to *section titles* of Wikipedia articles and construct sets of ''aspects'' through steps of automatic extraction, curation,

[1]http://github.com/neulab/wikiasp.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 211–225, 2021. https://doi.org/10.1162/tacl_a_00362
Action Editor: Asli Celikyilmaz. Submission batch: 3/2020; Revision batch: 8/2020; Published 3/2021.
© 2021 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

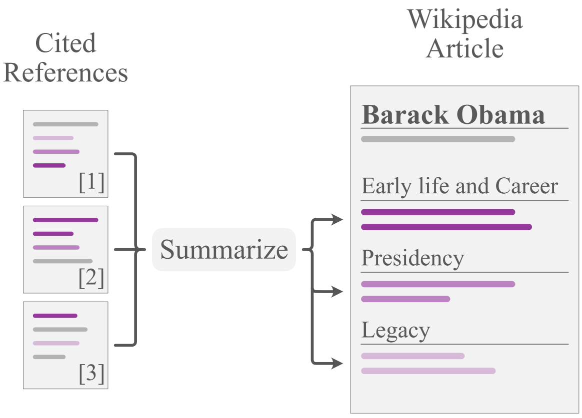

Figure 1: In WikiAsp, given reference documents cited by a target article, a summarization model must produce targeted aspect-based summaries that correspond to sections.

and filtering. The section texts then serve as corresponding aspect-based summaries.

We devise a baseline two-stage method consisting of aspect identification and summarization using extractive and abstractive models, and conduct experiments on the proposed dataset. The analysis of experimental results and the generated summaries reveals the unique challenges posed by our multi-domain and multi-document setting. For example, aspects that require summarizing contents in a particular order (e.g., time series events) in a multi-document setting adds extra difficulty because of the need for correctly ordering scattered (and possibly duplicate) pieces of information from different sources. Certain domains that involve interviews or quotes of people also exhibit challenges in correctly modifying pronouns based on the relationship to the topic of interest.

2 Generating Wikipedia as Aspect-based Summarization

Wikipedia articles exhibit a specific way of organizing information about a focused topic. An article S consists of two parts: section titles a, and their contents p. The contents are further split into sections, where each section describes information about the main topic from different viewpoints. Table 1 shows an example article about the topic ‘‘Barack Obama’’, with several sections ‘‘Early life and career’’, ‘‘Presidency’’, and ‘‘Legacy’’. In practice, the contents included in each section can take many forms, from text, tables, and images, to more specialized content

Title: Barack Obama
Aspect: *Early life and career*
Obama was born on August 4, 1961, at Kapiolani Medical Center for Women and Children in Honolulu, Hawaii. . . .
Aspect: *Presidency*
The inauguration of Barack Obama as the 44th President took place on January 20, 2009. In his first few days in office, Obama issued . . .
Aspect: *Legacy*
Obama’s most significant legacy is generally considered to be the Patient Protection and Affordable Care Act (PPACA), . . .

Table 1: Example Wikipedia article about Barack Obama. Our goal is to generate texts given the cited references and the specified aspects.

such as brackets of a tournament. In this work, we focus only on sections that mainly consist of textual content (see Section 3 for how we define this).

Importantly, the content in Wikipedia articles is required to be *verifiable*: ‘‘other people using the encyclopedia can check that the information comes from a reliable source’’.[2] To ensure this, articles contain citations from a set of references $\mathcal{R}$ so that readers can check the validity of the content. In other words, citations supposedly contain the majority of the information written in the articles. Liu et al. (2018) took advantage of this fact by proposing a summarization task using cited references as source documents for summarization. Citations include published material (such as books) and Web sites, but because only Web-based citations can easily and automatically be mined via crawling, we consider only Web-based citations as source documents in this work and ignore the rest of non-Web based citations following Liu et al. (2018).

The goal of our task is to learn a model $f : \mathcal{R} \rightarrow S$, which can 1) identify and gather information from cited references and 2) generate a section-by-section summary where each section contains the appropriate type of information. Formally, let $\mathcal{R} = \{R_1, R_2, \ldots, R_M\}$ be a collection of M cited references for an article $S = \{s_1, s_2, \ldots, s_N\}$ of N sections. Each section s_i is essentially a tuple of a section title and one or more paragraphs: $s_i = \langle a_i, p_i \rangle$.

[2] `https://en.wikipedia.org/wiki/Wikipedia:Verifiability`.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Although there is a fair amount of variety in section titles across different articles, articles that belong to the same domain tend to share aspects that are particularly salient for that domain. Because of this, we select a fixed-size subset of all section titles that appear in each domain as the set of aspects $\mathcal{A}$ that we will target; details on how we select this subset will be elucidated in the following section. Hence, our task is cast as multi-document aspect-based summarization.

3 The WikiAsp Dataset

In this section, we describe our concrete steps to create our dataset.

3.1 Data Collection

As the base data, we build upon the data collection strategy from the WikiSum dataset (Liu et al., 2018), a dataset for generating lead sections of Wikipedia from referenced Web pages. Following the WikiSum data generation script,[3] we first crawled cited references covered by CommonCrawl for each Wikipedia article. We then recover all the sections[4] of the target Wikipedia articles from the WikiSum (which was unused in the WikiSum dataset) and obtain pairs of (section title, section paragraph). An example for this is shown in Table 1.

3.2 Domain Separation

Articles in different domains focus on different salient topics, as observed by Perez-Beltrachini et al. (2019). For example, the ‘‘discography’’ section is common for articles about singers, but is not appropriate for articles about infrastructure. To characterize such structural differences, we separate the set of articles obtained in the previous step into sets in particular domains. Specifically, we follow Perez-Beltrachini et al. (2019) in assigning one category for each article using DBPedia (Auer et al., 2007). DBPedia stores structured information for each Wikipedia article, including the domain labels and info boxes. Additionally, it defines a topical hierarchy of the domains (ontology classes). We first map between articles and the domain labels from the corresponding DBPedia dump. Obtained domain labels, however, have mixed granularity (e.g., Person and its sub-class Dancer), which causes imbalance in the number of examples in each domain, as well as domain overlap between high-level and low-level domains in the domain hierarchy. We mitigate this by recursively merging domains at leaf-level into coarser ones according to the aforementioned topical hierarchy from the ontology classes.[5] We repeat the merging procedure until a branch in the hierarchy includes more than 15,000 articles, and picked 20 domains at the leaf of the merged hierarchy.[6]

3.3 Aspect Selection

Next, we perform aspect selection on each set of articles in the domains extracted during the previous step. As previously noted, articles in the same domain tend to share similar set of section titles. Motivated by this observation, we construct the set of aspects from the **most frequent section titles**.

From the frequency distribution of section titles in a domain, we manually filter ones that are not *textual*, that is, more than half portion of section consists of text. For each section title, we take 20 randomly sampled sections and include it in the set of aspects only if 80% of samples consist of *textual* paragraphs. Following the steps above, we construct the 10 most frequent aspects for each domain. However, the choice of words in section titles vary depending on the editors within the same domain, which leads to missing relevant aspects that are moderately frequent but not present in Top-10. For example, one of the common section titles in WrittenWork domain are ‘‘summary’’ and ‘‘plot summary,’’ which should be merged together to form a single aspect. We handle these cases by inspecting the frequent distribution further down and manually identifying semantically equivalent titles to merge.

The resulting dataset consists of instances in 20 domains where each domain has 10 pre-defined aspect classes. We show statistics comparisons of the dataset to existing aspect-based summarization

[3] Tensor2tensor’s WikiSum generator was used.

[4] Due to the design of WikiSum dataset, the first section title of any article is automatically renamed to ‘‘LEAD’’. Therefore, we could not recover first sections of the Wikipedia articles. We suggest editing the data generation scripts for future WikiSum users if section title information is necessary.

[5] `http://mappings.dbpedia.org/server/ontology/classes/`.

[6] Many articles are labeled directly as Person, in which case the domain is high-level at the hierarchy. We do not select this domain because lower-level domains such as Artist or SoccerPlayer already have enough articles.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Infrastructure		Software	
history	13293	reception	8196
route description	5627	gameplay	8095
facilities	2792	development	3983
services	1955	plot	3697
future	784	history	2465
route	689	features	1799
location	613	story	991
construction	577	release	750
connections	497	overview	570
description	463	legacy	564

Table 2: Frequency of filtered aspects that are *textual* in 2 domains. Due to space constraint, the statistics for the rest of domains will be available in the Appendix C.

datasets in Table 3 and examples of obtained aspects for two domains in Table 2.

Appendix A and C summarizes the data size for each domain and the obtained aspects for the rest of 18 domains respectively.

4 Baseline Models

Next, in this section we describe two baseline models for solving this task. Both of these models decompose the overall process into two stages: aspect discovery and aspect-based summarization of classified sentences. Both baseline models share the same methodology for aspect discovery, but differ in terms of summarization models. The model overview is shown in Figure 2.

4.1 Aspect Discovery

The first stage consists of labeling sentences in *cited reference texts* according to aspects. Having training data that contains sentences in the reference documents labeled with target aspects would be the ideal case, but these do not exist a priori. Therefore, we instead create training data by assigning each sentence in the target articles with aspect labels corresponding to the aspect to which the sentence belongs. For example, the article about Barack Obama in Table 1 yields training instances consisting of sentences labeled with *Early life and career*, *Presidency*, and *Legacy* depending on which paragraph a sentence comes from. This data makes it possible to train a classifier that learns to predict aspects from the texts at sentence-level. At test time, cited reference sentences are fed into the learned classifier and are labeled with their most likely aspects.

However, the discrepancy of inputs at train/test time is problematic because the model is not exposed to any *noisy* sentences that do not belong to any of the relevant aspects at training time, while cited reference texts do contain such sentences. For example, an article in the *Company* domain may have a citation to the company Web site itself, which contains commercial messages that may not be appropriate in encyclopedic text such as Wikipedia. We manage such cases by introducing an auxiliary label *Other* at training time and let the model learn to identify noisy sentences as well. To do so, sentences labeled with *Other* are randomly sampled from texts in different domains and added to training data. We fine-tune the pretrained ROBERTa (Liu et al., 2019) model on this classification dataset for each domain. Logits obtained from the model are then passed through the sigmoid function to obtain probabilities of each aspect for a given sentence. Finally, we assign labels to a sentence by taking the aspects a_i whose probabilities are greater than the threshold λ: $P(a_i) > \lambda$. The lower we set the threshold, the more but potentially noisy sentences we include as the input to the summarization model. We tune λ independently for each domain based on the performance on validation sets and set 0.5 for *Group*, 0.8 for *Album*, *Animal*, *Building*, *Film*, and 0.9 for the remaining domains as the threshold values.

4.2 Summarization

Sentences that are labeled with the same aspect are then grouped in order of occurrence in cited references to form a chunked paragraph that discusses the same aspect. This forms aspect-based clusters of relevant sentences, which become the input to a summarization model. On the contrary, aspects that are never labeled (due to low probabilities) are deemed irrelevant and thus will not be summarized. We consider both an extractive and an abstractive summarization model in our baseline implementation. For the extractive model, we use TextRank (Mihalcea and Tarau, 2004; Barrios et al., 2016), a graph-based ranking model for extracting important sentences. For the abstractive model, we use PreSumm (Liu and Lapata, 2019b), a Transformer-based summarizer with fine-tuned BERT as the source encoder. For each domain, PreSumm is fine-tuned and trained on the pairs of (grouped sentences, target aspect

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Dataset	Domain	#Dom.	#Train	Doc. Length	Sum. Length	#Asp.	#Asp./Ex.
OpoSum	Product Review	6	359,048	138	49	9	2.00
Amazon	Product Review	7	240,000	82	–	–	–
RottenTomatoes	Movie Review	1	2,458	2369	24	*2	*1.00
MA-News	News	1	284,701	1350	54	6	2.98
WikiAsp	Encyclopedia	20	320,272	13,672	213	10	1.77

Table 3: Training set statistics comparisons against previous aspect-based summarization datasets. For multi-domain datasets, the sum of all the examples are reported. #Asp./Ex. represents the average number of aspects that a model has to summarize on each example. (*Review saliency is treated as aspects. #Asp. represents the number of aspects per domain if the number of domains is more than one. Compared datasets are the work of Angelidis and Lapata (2018); Yang et al. (2018); Wang and Ling (2016); Fermann and Klementiev (2019), respectively.

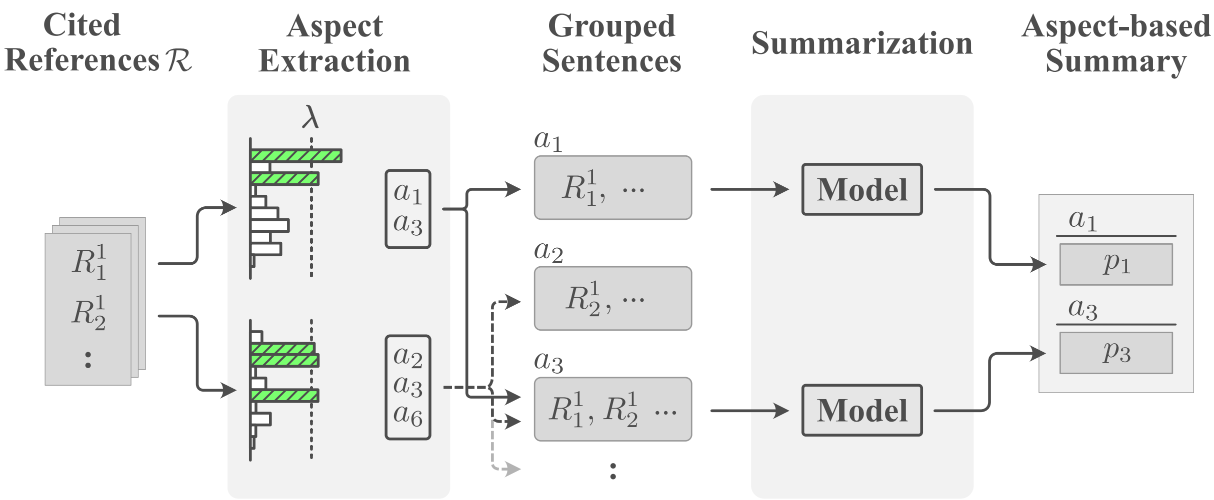

Figure 2: Two-stage model diagram. The aspect classifier assigns aspect labels for each reference sentence R_j^i from references $\mathcal{R}$ with a threshold λ. Sentences are then grouped according to the assigned labels, which are fed to the summarization model. Groups about irrelevant aspects (i.e., a_2) is ignored. Finally, the summarization model outputs summaries for each relevant aspect.

paragraph) to learn to produce summaries given the aspect-relevant sentences.

5 Evaluation

We evaluate models along two axes: aspect discovery and summarization. We note that the primary task in this dataset is aspect-based summarization, thus aspect discovery evaluation discussed below is only for diagnostic purposes. Because the aspect sets differ in different domains, evaluation is performed separately for each domain.

Aspect Discovery Models have to correctly predict the right set of aspects about which they generate summaries. The aspect discovery criterion aims to evaluate the similarity between the set of aspects about which a model decides to generate summaries and the set of aspects that appear in the target article.[7] For comparing these two sets, we use precision, recall, and F1 scores.

Aspect-based Summarization Gold standard summaries only exist for each of the aspects that appear in an article. Therefore in this evaluation, we focus on evaluating the model's ability to summarize inputs particularly on these aspects. Specifically, generated summaries are paired to

[7]Note that there are two potential reasons an aspect does not appear in the target article: (1) it may not be appropriate for that particular entity (e.g., the ''controversy'' aspect in the ''company'' domain should not exist if that company has legitimately never had a controversy), or (2) the article may not be complete. For this evaluation, we make the simplifying assumption that all articles are complete and thus missing aspects are an indication of failure to recall information, but relaxing this assumption in some way may result in more accurate evaluation.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

corresponding reference summaries with the same aspects and are evaluated using ROUGE (Lin, 2004). Because ROUGE is a recall-based measure, the number of tokens in the model outputs directly affect the performance. Controlling the length is particularly important for our dataset because average summary length for each aspect in different domains varies (e.g., ''description'' and ''location'' from HistoricPlace domain has 396 and 90 average tokens, respectively). We take this into account by explicitly setting the maximum number of words for extractive and abstractive summaries to be the average number of words in the target summaries in the training set for each aspect and for each domain.

6 Experiments

We provide two baseline models for the task and evaluate on the proposed dataset.

6.1 Implementation Details

For aspect classification, we used the `roberta-base`[8] model and fine-tuned for 5 epochs on the created surrogate dataset above for each domain, with the learning rate 2×10^{-5}. For the extractive summarization, we specify the summary length for TextRank according to the mean length of target summaries for each aspect in each domain. We re-train the PreSumm summarizer on our dataset for each domain: the encoder is initialized with the weights of pre-trained BERT (Devlin et al., 2019) and the decoder is trained from scratch. The total number of training steps is 300,000. For some domains, we further tuned the decoder dropout rate to 0.3 to stabilize training. At inference time, we specify maximum summary lengths for each *aspect* for each domain using the average summary lengths from computed from the training set.

6.2 Results

In this section, we discuss the experimental results at each stage.

6.2.1 Aspect Discovery

We show the aspect discovery results in Table 4. We see a general trend of high recall predictions made by the model. While varying thresholds could balance precision and recall, the results exhibited high recall after hyperparameter search.

Domain	Prec	Rec	F-1
Album	19.64	86.43	30.64
Animal	34.69	84.08	45.52
Artist	26.32	75.24	36.72
Building	31.46	91.25	42.92
Company	28.97	91.50	41.06
EducationalInstitution	25.64	93.82	37.66
Event	28.99	96.44	42.36
Film	32.84	91.46	45.17
Group	17.46	95.56	28.18
HistoricPlace	33.38	90.22	42.98
Infrastructure	28.38	94.00	41.00
MeanOfTransportation	23.24	83.13	33.88
OfficeHolder	21.22	73.25	30.62
Plant	31.25	83.17	42.10
Single	25.36	88.33	37.16
SoccerPlayer	28.54	67.18	37.16
Software	31.52	94.65	45.10
TelevisionShow	20.44	81.76	31.28
Town	42.61	71.85	50.12
WrittenWork	21.50	94.29	33.71

Table 4: Aspect discovery results on the test set.

This suggests that the learned classifier is poorly calibrated. Class imbalance also plays a role here; predicting the major classes give high recall due to skew aspect frequency distributions. Among others, the classifier performed best with the Town domain by achieving the highest precision and the F1 score.

6.2.2 Summarization

The automatic evaluation results are shown in Table 5. Neither baseline unanimously outperformed the other on all domains, but we observe that PreSumm (abstractive) performs better than TextRank (extractive) on average. The low R-2 and R-L scores by both models despite the oracle being relatively higher suggest that important phrases to be summarized do not appear rarely.[9]

To understand the upper-bound of model performance for the task, we also show summarization results of the extractive oracle model in Table 5. Sentences were chosen directly from cited reference texts to maximize the ROUGE score against summaries, thus bypassing the aspect classification stage. The oracle performance shows that a summarization model can indeed perform competitively on the dataset if the model is given with the full input information. The contrasting results

[8]We used Huggingface's implementation (Wolf et al., 2019) for obtaining and fine-tuning the weights.

[9]Note that TextRank connects nodes according to content overlap, thus isolated sentences are not selected.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

	TextRank			PreSumm			Extractive Oracle		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Album	19.56	2.81	17.26	22.76	6.31	20.27	37.72	12.58	33.19
Animal	18.00	3.16	16.05	27.11	8.08	25.01	34.82	10.52	31.01
Artist	17.22	2.49	15.58	21.79	3.76	20.00	41.49	15.04	37.64
Building	23.91	4.96	21.85	24.99	5.97	23.24	41.95	14.31	38.28
Company	22.92	3.70	20.65	22.28	4.08	20.50	40.20	12.30	36.16
EducationalInstitution	21.47	4.29	19.24	24.17	6.70	21.96	39.11	14.04	35.18
Event	26.64	5.67	24.08	28.31	7.69	26.20	46.17	16.90	41.87
Film	21.25	3.81	19.14	20.58	5.34	18.86	40.24	13.78	36.14
Group	22.30	3.62	20.20	25.51	4.97	23.51	41.36	13.23	37.56
HistoricPlace	18.96	3.71	17.51	27.40	8.08	25.69	37.78	10.83	34.65
Infrastructure	20.40	3.27	18.39	27.86	9.24	25.80	36.04	10.00	32.25
MeanOfTransportation	21.20	3.93	19.31	24.52	7.04	22.72	41.13	13.70	37.45
OfficeHolder	18.45	3.15	16.77	19.63	5.24	18.12	39.60	14.70	36.04
Plant	18.73	3.02	16.84	25.29	6.30	23.20	34.93	9.66	31.31
Single	17.96	2.67	15.86	22.06	6.78	19.98	36.51	11.57	31.88
SoccerPlayer	14.79	2.36	12.89	12.89	1.86	12.05	31.06	8.00	27.08
Software	24.54	4.56	22.05	20.51	5.15	18.82	42.79	13.96	38.30
TelevisionShow	19.77	3.21	17.68	19.20	3.53	17.42	40.35	13.47	35.67
Town	17.89	3.56	16.50	19.76	4.39	16.87	33.21	10.31	30.70
WrittenWork	23.39	3.89	21.14	22.19	4.33	20.15	42.66	13.93	38.16
AVG	20.47	3.59	18.45	22.94	5.74	21.02	38.95	12.64	35.03

Table 5: Aspect-based summarization results on the test set. The last row shows the average performance.

between the oracle and two stage models suggests the importance of accurate content selection before performing summarization.

7 Analysis

We discuss the model outputs and analysis below.

7.1 Aspect-by-Aspect Evaluation

Not all the aspects are equally hard to summarize; some might require summarization of a broad range of information, whereas others require only specific concepts to be summarized. We further investigate this by looking into summarization performance for both models on per-aspect basis. Table 6 shows the best-performing aspects sorted in descending order by ROUGE-1 scores for two summarization models on the validation set. Through manual investigation of the generated samples for each aspect, we observed that the aspects where the *abstractive* model performed well tend to have common templates and similar choice of vocabulary, more so than other aspects. For example, 58% (out of 183 samples) of the target summaries for *government* in Town shared the identical summaries despite the fact that articles discuss different townships. Similar but less prevalent patterns were observed in other aspects as well.

Aspects where the *extractive* summarization model performed better contain much larger numbers of tokens in the summaries than average. Specifically, the average summary length for 10 aspects where TextRank performed the best was 303, while that for 10 aspects where PreSumm performed the best was 166. Naturally, abstractive models have issues with maintaining coherence over long decoding results, but the extractive model has few issues gathering relevant sentences at the cost of incoherent transitions from sentence to sentence. As for the content, extractive summaries exhibited the advantage of being able to correctly include mentions related to numbers and dates.

7.2 Quality of Generated Summaries

We then examined the generated summaries from the two models and compared them qualitatively. Samples are shown[10] in Table 7 from some of the domains listed in Table 2.

Manual inspection of the generated summaries revealed pros and cons of the two models:

[10]Samples from other domains are in Appendix B.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Dom.	Aspect	PreSumm	TextRank
		↓ R-1	R-1
Tow.	government	**55.10**	21.20
Eve.	format	44.94	24.73
Inf.	facilities	42.46	14.75
Bui.	exterior	41.81	25.60
Mea.	background	39.00	23.72
His.	heritage listing	36.58	10.25
Ani.	habitat	32.91	12.95
Pla.	taxonomy and nm.	32.70	9.39
Edu.	rankings	31.80	26.92
Alb.	commercial perf.	31.71	15.51
Dom.	Aspect	R-1	↓ R-1
Eve.	battle	28.00	**32.00**
Eve.	report	24.77	30.11
Sof.	gameplay	24.17	28.53
Eve.	background	30.01	27.42
Eve.	aftermath	27.54	27.27
Bui.	history	25.32	27.13
Sof.	plot	20.50	27.00
Edu.	rankings	31.80	26.92
Wri.	plot summary	22.08	26.85
Fil.	plot	19.43	26.66

Table 6: List of aspects sorted in descending order of ROUGE-1 score according to PreSumm (top half) and TextRank (bottom half). ‘‘performance’’ and ‘‘naming’’ are abbreviated to ‘‘perf.’’ and ‘‘nm.’’, respectively. Domain names shortened to the first three letters.

- **Both models are successful at discussing on-topic content.** For all the summaries inspected, both models were able to generate on-topic content in spite of the source documents potentially being noisy.
- **Abstractive summaries underperform at generating exact entity mentions.** Almost all the samples require generation of entities because the task targets at generating encyclopedic texts. Except for the title (topic) entity, abstractive models either generated no entities or wrong ones.

7.3 Aspect Classification Accuracy

We observed a general trend of low precision for aspect discovery. We hypothesize that this is due to limited target aspects for each article; correctly extracted aspects affect negatively to precision if they do not exist in the target article. To quantify this, 10 random articles are selected from the validation set in Software domain. For each article, we extract 10 sentences labeled with the highest confidence for each of the 10 aspects, resulting in 1,000 sentences in total. Each sentence is annotated with binary labels indicating whether it is correctly associated with the aspect or not.[11] With the threshold λ set to 0.9, we achieved the precision of 45.1, which shows that the aspect discovery has the ability to extract aspects, but is not as good at extracting *relevant* aspects for the article. We observed that the model predictions tend to be polarized to extreme values (i.e., near 0 or 1). We also show the relationship between λ ranges and the precision in Figure 3, which indicates that the classifier is not well calibrated.

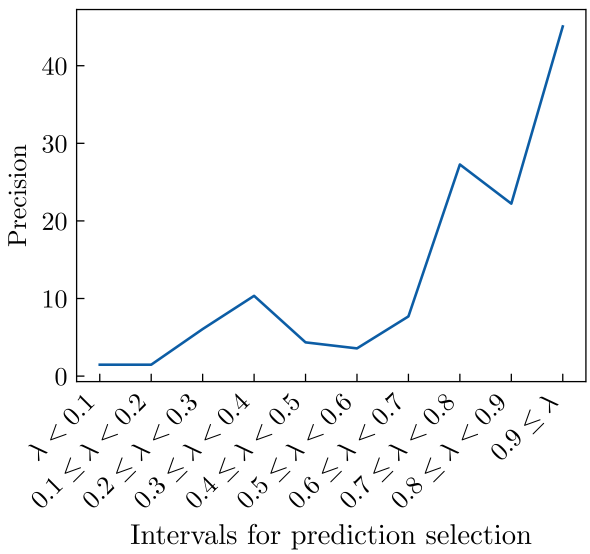

Figure 3: Precision differences in varying threshold ranges.

7.4 Domain-specific Challenges

One of the benefits of having many domains for the same task is to be able to characterize the differences and challenges that are unique to certain domains. We analyzed the generated summaries from both of the summarization models and identified some of them below.

7.4.1 Pronoun Resolution for Opinion-based Inputs

This is particularly important in domains and aspects with subjective reviews such as Music (Album, Artist, Group, and Single) or Software. Source documents in these domains often include quotes by artists or critics, which are often written from different person perspective. These are

[11] Sometimes, the entity in discussion by the sentence is not clear. In this case, we annotate it correct if the sentence could correspond to the target aspect of any entity.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Domain / Title: Software / Cyberpunk 2077	
Aspect: *Gameplay*	
Gold: cyberpunk 2077 is a role - playing video game played from either a first - person or third - person perspective . it is set in an open world metropolis called night city . the game will feature non - english speaking characters . players who do not speak the languages can buy translator implants to better comprehend them; . . . **Ext.**: cyberpunk 2077 takes place in, you guessed it, the year 2077 . for just a few hours, you can be rich, successful, and popular with your preferred gender . cyberpunk 2077 will be a aaa rpg, but whether it will come to resemble the witcher in any way remains unclear . how braindances will be used by or on the protagonist is . . . **Abs.**: the game is an action role - playing game in which the player assumes the role of an unnamed protagonist, runabout who lives in a large metropolis known as the ‘‘black city’’ . the game’s narrative is told through procedurally generated procedurally generated . during the course of the game, the player is briefed into . . .	
Domain / Title: Event / Battle of the Saintes	
Aspect: *Aftermath*	
Gold: the british lost 243 killed and 816 wounded, and two captains out of 36 were killed . the french loss in killed and wounded has never been stated, but six captains out of 30 were killed . it is estimated that the french may have lost as much as 3,000, and more than 5,000 french soldiers and sailors were captured, . . . **Ext.**: battle of the saintes, (april 9 – 12, 1782), in the american revolution, major naval victory for britain in the west indies that restored british naval mastery in the area and ended the french threat to nearby british possessions . the valiant was subsequently present at the relief of gibraltar on 12 april 1781 . . . **Abs.**: the immediate aftermath of the battle of the niles meant the british fleet could not be able to deliver a great strategic victory in the kingdom of france, and had also declared war on great britain with the royal navy to france . the outcome was the subject of a strategic victory for the british . in the meantime, the british navy had captured the french navy . . .	
Aspect: *Battle*	
Gold: on 9 april 1782, the copper - hulled british fleet soon caught up with the french, who were surprised by their speed . de grasse ordered the french convoy to head into guadeloupe for repair, forcing him to escort two fifty - gun ships (fier and experiment) and placing his fleet in line of battle in order to cover the retreat . . . . **Ext.**: ‘‘after some initial maneuvers and minor clashes, a full - scale battle was joined on april 12, by which time the british had thirty - six ships of the line in action against thirty french ones . he turned his ships ninety degrees and sailed through the broken french line of battle, splitting the french line into four segments . in doing this, the guns on each side of the british ships were brought to bear on the french with little risk of return fire . . . . **Abs.**: the battle started as the shock . the battle progressed, when the british forces reached the north - eastern flank of the battle of weidman (kingdom) in a battle . he had begun to return to the field, and moved his forces toward the mouth of the river . in the battle, the first contingent of the french navy ships got off from a small contingent of british soldiers as well as the third - rate, under the command of general sir henry sturgis	

Table 7: Generated summaries from multiple domains. Ext. and Abs. represent summaries from TextRank and PreSumm.

usually converted by the Wikipedia editors into more encyclopedic text, citing the source of the information and writing in the third person. By design, extractive summaries have issues with this problem because of the lack of ability to transform the input sentences in any way. For example, the first extractive summary in Table 7 describes a game in a subjective way. We verified this by randomly selecting 20 summaries for *gameplay* aspect in Software domain. We inspected pronouns in extractive summaries and mark ones with first- or second-person pronouns if the gold summaries do not contain them. We found 65% of the samples contained those undesirable pronouns that do not align with the format of gold summaries.

7.4.2 Chronological Explanation

This variety of content is often found in certain aspects such as *history* and *event*, which tend to appear across multiple domains but are most prevalent in Event, HistoricPlace, and non-human entities like Company and Building. It is essential in these aspects to describe key information in the right chronological order for better readability. This would not be a hard task for single document summarization, as the model could perform reasonably by following the order of the original document. However, because our input is of multi-document form, maintaining chronological order when aggregating information across multiple domains becomes non-trivial. Indeed, neither of the models were successful at being truthful to the order even when there are enough clues in the original references. For example, multiple sentences start with ‘‘In [year], . . .’’, but the generated summary jumps around in time. We randomly picked 20 samples of extractive summaries with *history* aspect from Company domain and found that 25% of the samples have inconsistent timeline explanations.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

8 Related Work

Aspect-based Summarization

Aspect-based summarization has been widely investigated primarily on product or restaurant reviews (Titov and McDonald, 2008; Lu et al., 2009; Yang et al., 2018; Wang and Ling, 2016). Angelidis and Lapata (2018) proposed a weakly supervised method for aspect-based opinion summarization that discovers aspects with a topic model and does not require gold aspect annotation. TAC 2010 held a shared task of guided-based summarization on newswire domain, which resembles aspect-based summarization in terms of topic guidance. Recently, the task has been extend to news-domain by generating artificial datasets for aspect-based summarization to address the lack of large-scale data with aspect annotation (Frermann and Klementiev, 2019; Krishna and Srinivasan, 2018). Our work also builds an aspect-based summarization dataset automatically and is most similar to Krishna and Srinivasan (2018), but utilizes naturally available online encyclopedia entries and their sections in multiple domains.

Wikipedia as a Summarization Dataset

Wikipedia has been studied as a target resource for generation. An early attempt on generating *full* Wikipedia articles relied on Web search results for target entities as inputs (Sauper and Barzilay, 2009), which simulates an authoring process of humans searching information over the Internet. Liu et al. (2018) formulate a sub-task of generating *lead* sections as summarization of reference web pages to target articles. The resulting WikiSum dataset is accompanied by rich metadata about articles and inspired different uses of the dataset (Perez-Beltrachini et al., 2019). Our work also builds upon the WikiSum dataset, and aims to evaluate aspect-based summarization models using different sections from Wikipedia articles. Compared with Sauper and Barzilay (2009), our dataset is an order of magnitude larger, both in the number of articles and in the number of domains covered.

Multi-document Summarization

Extractive methods have shown effective for multi-document summarization in previous work (Nenkova et al., 2006; Cao et al., 2015; Yasunaga et al., 2017), but abstractive methods have increasingly adopted for the task (Lebanoff et al., 2018; Fabbri et al., 2019). Our task is based on the idea of (Liu et al., 2018) which treats references as source documents for the multi-document summarization task, and we experimented with both types of summarization models in our experiments.

9 Conclusion and Future Work

In this paper, we propose a large-scale, multi-domain multi-aspect summarization dataset derived from Wikipedia. Through experiments, we perform an extensive analysis of performance across different genres and aspect types. Our analysis has demonstrated that there are both general challenges regarding summarization into various aspects, as well as specific challenges in particular genres such as time-consistent mentions and proper pronoun conversion depending on the writer of the original content.

Because of this, the proposed datas et also provides a testbed for several potential directions for future work. For example, better aspect discovery models may take into account the coherence of the discourse in the original documents when extracting aspects. Better summarization models may take into account the provenance of the information, appropriately determining when the information is written by a first or third party. WikiAsp also invites a focus on domains of interest to investigate various problems of text summarization, such as correct pronoun handling and description of chronological timeline.

Acknowledgments

We would like to thank anonymous reviewers for insightful comments. HH and GN were supported by a grant from AlphaSense.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1403

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

open data. In *The semantic web*, pages 722–735. Springer. **DOI:** https://doi.org/10.1007/978-3-540-76298-0_52

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Twenty-ninth AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1102

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. 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)*, pages 4186–4196, Hong Kong, China. Association for Computational Linguistics.

Lea Frermann and Alexandre Klementiev. 2019. Inducing Document Structure for Aspect-based Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1630

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/v1/N15-1113

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N18-1065

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661. New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N18-1149

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1208

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating Topic-Oriented Summaries Using Neural Attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1697–1705, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N18-1153

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1446

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. *arXiv:1801.10198 [cs]*. ICLR.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1500

Yang Liu and Mirella Lapata. 2019b. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1387

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv: 1907.11692 [cs]*.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web - WWW '09*, page 131, Madrid, Spain. ACM Press. **DOI:** https://doi.org/10.1145/1526709.1526728

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/K16-1028

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 573–580. ACM. **DOI:** https://doi.org/10.1145/1148170.1148269

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating summaries with topic templates and structured convolutional decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5107–5116, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1504

Christina Sauper and Regina Barzilay. 2009. Automatically Generating wikipedia articles: a structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216,

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Suntec, Singapore. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1687878.1687909

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N16-1007

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. 2018. Aspect and sentiment aware abstractive review summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1110–1120, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/K17-1045

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

A Domain Statistics

Domain	Train	Valid	Test
Album	24434	3104	3038
Animal	16540	2005	2007
Artist	26754	3194	3329
Building	20449	2607	2482
Company	24353	2946	3029
EducationalInstitution	17634	2141	2267
Event	6475	807	828
Film	32129	4014	3981
Group	11966	1462	1444
HistoricPlace	4919	601	600
Infrastructure	17226	1984	2091
MeanOfTransportation	9277	1215	1170
OfficeHolder	18177	2218	2333
Plant	6107	786	774
Single	14217	1734	1712
SoccerPlayer	17599	2150	2280
Software	13516	1637	1638
TelevisionShow	8717	1128	1072
Town	14818	1911	1831
WrittenWork	15065	1843	1931

Table 8: The list of domains and the number of Wikipedia articles in each domain that contain at least one salient aspect.

B Additional Samples

Title: Recomposed by Max Richter: Vivaldi – The Four Seasons

Aspect: *Critical Reception*

Gold: recomposed by max richter: vivaldi - the four seasons received widespread acclaim from contemporary classical music critics . ivan hewett of the telegraph gave the album a very positive review, stating, " as you would expect of a composer who once studied with the great modernist luciano berio, richter is very self - aware . . . .

Ext.: listen to recomposed by max richter: vivaldi, the four seasons now . i am highly impressed with 'recomposed'. the music then propels the audience into an atmosphere of isolation; a delicate harmony that is sustained whilst hope takes centre stage . . . .

Abs.: the allmusic review by michael g . nastos awarded the album 4 stars stating ‘ ‘ this is an album that generally considered for fans of the genre ‘ ‘ . . . .

Table 9: Generated summaries from **Album** domain.

Title: Pride and Glory (film)

Aspect: *Plot*

Gold: assistant chief francis tierney sr . is the head of a multigenerational new york city police department (nypd) family, which includes his sons francis "franny" jr . , ray, and his son - in - law jimmy egan . deputy inspector franny is the commanding officer of the 31st precinct, where sergeant jimmy is a patrol officer, . . .

Ext.: as we know, under the macho code, this means that after two people who love each other end up beaten and bloody, they will somehow arrive at a catharsis . the plot involves how and why the four cops were killed . a family of police officers - patriarch, two sons, and a son - in - law - deals with corruption in a precinct in washington heights . . . .

Abs.: in the year before the events of the first film, the movie takes place in washington heights, d . c . , a . army sergeant - in - law, ray ' s wife, and sister abby, living in washington city . they have a romantic relationship with one of their officers . while the four officers are called to ‘ ‘ the mental patient ‘ ‘ , . . .

Table 10: Generated summaries from **Film** domain.

Title: Dimitri Soudas

Aspect: *Career*

Gold: soudas served for one term as a school trustee at the western quebec school board from 2002 to 2005 . between 2006 and 2011, soudas was a "high profile" member of prime minister stephen harper's communication team, and one of the prime minister's "closest and most faithful aides" initially serving as a press secretary and later as an associate director of communications for the prime minister ' s office, . . .

Ext.: april 2010 – after serving as a press secretary in the prime minister's office, soudas was promoted to director of communications . "to fulfil the opportunities afforded by social media, directors of communication need to be aware of this trend and engage with it," dimitri soudas writes in his master's thesis, a copy of which has been obtained by cbc news. . . .

Abs.: in 2001, he was elected to the canadian house of commons as a member of the people's action party (pc) for the riding of yorkshire . he was re - elected in 2002 and 2006 . in 2006, he was .

Table 11: Generated summaries from **Office-Holder** domain.

C Aspect Statistics

Tables 12 and 13 show aspect frequency statistics. Perf., hist., dist., ext., desc., dev., edu., nm., and intl. correspond to performance, history, distribution, extracurricular, description, development, education, naming, and international, respectively.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023

Album		Animal	
reception	11782	description	12729
critical reception	6682	distribution	7813
background	6202	dist. & habitat	2967
commercial perf.	2398	taxonomy	2737
release	2209	habitat	2208
chart positions	1891	behavior	2167
recording	1490	ecology	1777
promotion	1150	diet	1363
history	1045	reproduction	1291
overview	840	biology	1238
Artist		**Building**	
career	10193	history	16885
biography	8292	architecture	3223
early life	7587	desc. & hist.	1395
personal life	6775	description	1382
music career	2829	location	906
death	1607	interior	877
life and career	1512	construction	862
early life & edu.	1239	exterior	746
early years	1129	design	623
exhibitions	1030	facilities	572
Company		**EducationalInstitution**	
history	21488	history	12798
products	2921	athletics	5602
operations	1630	academics	4638
services	1019	campus	2471
controversy	920	sports	1433
overview	891	student life	1327
background	572	ext. activities	1227
subsidiaries	556	curriculum	1191
company history	504	facilities	1189
technology	471	rankings	836
Event		**Film**	
background	3453	plot	25772
aftermath	2483	reception	14003
history	1361	production	13882
battle	1228	release	7299
format	461	box office	4572
prelude	450	critical reception	4195
event	416	critical response	2802
report	323	synopsis	2626
summary	321	home media	2461
casualties	290	filming	2013

Table 12: Aspect frequency for 8 domains.

Group		HistoricPlace	
history	8894	history	3232
biography	1206	description	1398
career	1102	desc. & hist.	1250
musical style	683	heritage listing	942
background	581	architecture	549
formation	408	location	161
early years	279	historic uses	90
legacy	272	preservation	84
style	265	geography	75
influences	204	interior	70
MeanOfTransportation		**OfficeHolder**	
history	2572	personal life	5119
design	2152	political career	4950
operational hist.	1989	early life	4740
design & dev.	1566	career	4115
service history	1435	biography	2801
development	1096	education	2168
construction	933	background	1578
fate	632	death	1402
background	604	legacy	889
description	602	early life & career	859
Plant		**Single**	
description	4684	music video	9606
dist. & habitat	1649	critical reception	3829
uses	1585	background	3459
distribution	1399	reception	2097
cultivation	1387	composition	1729
taxonomy	1121	cover versions	1594
ecology	884	content	1266
conservation	554	release	1045
etymology	389	commercial perf.	849
taxonomy & nm.	384	live performance	113
SoccerPlayer		**TelevisionShow**	
intl. career	8055	plot	2902
club career	8029	production	2648
career	6386	reception	2643
personal life	3621	synopsis	1304
playing career	1930	premise	944
early career	1578	history	908
early life	1191	format	842
professional	992	broadcast	779
style of play	887	overview	650
football career	550	critical reception	583
Town		**WrittenWork**	
geography	12667	plot	5495
demographics	10949	reception	4970
history	7298	plot summary	3900
education	2868	history	2527
government	1910	background	1218
2010 census	1363	adaptations	1173
2000 census	1284	critical reception	933
transportation	1239	manga	830
economy	1066	history and profile	803
name and history	1002	anime	714

Table 13: Aspect frequency for 10 domains.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00362/1924027/tacl_a_00362.pdf by guest on 09 December 2023