# *KPTimes*: A Large-Scale Dataset for Keyphrase Generation on News Documents

**Ygor Gallina** **Florian Boudin** **Beatrice Daille**

LS2N, Université de Nantes, France
firstname.lastname@univ-nantes.fr

## Abstract

Keyphrase generation is the task of predicting a set of lexical units that conveys the main content of a source text. Existing datasets for keyphrase generation are only readily available for the scholarly domain and include non-expert annotations. In this paper we present *KPTimes*, a large-scale dataset of news texts paired with editor-curated keyphrases. Exploring the dataset, we show how editors tag documents, and how their annotations differ from those found in existing datasets. We also train and evaluate state-of-the-art neural keyphrase generation models on *KPTimes* to gain insights on how well they perform on the news domain. The dataset is available online at https://github.com/ygorg/KPTimes.

## 1 Introduction

Keyphrases are single or multi-word lexical units that best summarise a document (Evans and Zhai, 1996). As such, they are of great importance for indexing, categorising and browsing digital libraries (Witten et al., 2009). Yet, very few documents have keyphrases assigned, thus raising the need for automatic keyphrase generation systems. This task falls under the task of automatic keyphrase extraction which can also be the subtask of finding keyphrases that only appear in the input document. Generating keyphrases can be seen as a particular instantiation of text summarization, where the goal is not to produce a well-formed piece of text, but a coherent set of phrases that convey the most salient information. Those phrases may or may not appear in the document, the latter requiring some form of abstraction to be generated. State-of-the-art systems for this task rely on recurrent neural networks (Meng et al., 2017; Chen et al., 2018, 2019), and hence require large amounts of annotated training data to achieve good performance. As gold annotated data is expensive and difficult to obtain (Mao and Lu, 2017), previous works focused on readily available scientific abstracts and used author-assigned keyphrases as a proxy for expert annotations. However, this poses two major issues: 1) neural models for keyphrase generation do not generalize well across domains, thus limiting their use in practice; 2) author-assigned keyphrases exhibit strong consistency issues that negatively impacts the model's performance. There is therefore a great need for annotated data from different sources, that is both sufficiently large to support the training of neural-based models and that comprises gold-standard labels provided by experts. In this study, we address this need by providing *KPTimes*, a dataset made of 279 923 news articles that comes with editor-assigned keyphrases.

Online news are particularly relevant to keyphrase generation since they are a natural fit for faceted navigation (Tunkelang, 2009) or topic detection and tracking (Allan, 2012). Also, and not less importantly, they are available in large quantities and are sometimes accompanied by metadata containing human-assigned keyphrases initially intended for search engines. Here, we divert these annotations from their primary purpose, and use them as gold-standard labels to automatically build our dataset. More precisely, we collect data by crawling selected news websites and use heuristics to draw texts paired with gold keyphrases. We then explore the resulting dataset to better understand how editors tag documents, and how these expert annotations differ from author-assigned keyphrases found in scholarly documents. Finally, we analyse the performance of state-of-the-art keyphrase generation models and investigate their transferability to the news domain and the impact of domain shift.



*Proceedings of The 12th International Conference on Natural Language Generation*, pages 130–135,
Tokyo, Japan, 28 Oct - 1 Nov, 2019. 

**Muslim Women in Hijab Break Barriers: 'Take the Good With the Bad'**

When Ginella Massa, a Toronto-based TV reporter, recently accepted a request to host an evening newscast, she was not planning or expecting to make history for wearing a hijab. She was just covering for a colleague who wanted to go to a hockey game. And that's how Ms. Massa, who works at CityNews in Toronto, became the first Canadian woman to host a newscast from a large media company while wearing the head scarf. [...] This new trend of inclusion occurs amid a more sinister one, as reported hate crimes against Muslims are on the rise in the United States and Canada. The F.B.I. says that a surge in hate crimes against Muslims has led to an overall increase in hate crimes in the United States; Muslims have borne the brunt of the increase with 257 recorded attacks. [...] In Canada, where Ms. Massa has lived since she was a year old, the number of reported hate crimes has dropped slightly overall, but the number of recorded attacks against Muslims has grown: 99 attacks were reported in 2014, according to an analysis by the news site Global News of data from Statistics Canada, a government agency. [...]

**keywords:** US; Islam; Fashion; Muslim Veiling; Women and Girls; (News media, journalism); Hate crime; Canada

Figure 1: Sample document from *KPTimes* (id: 0296216). Keyphrases (or part of) appearing in the document are colored.

## 2 Existing datasets

Frequently used datasets for keyphrase generation have a common characteristic that they are, by and large, made from scholarly documents (abstracts or full texts) paired with non-expert (mostly from authors) annotations. Notable examples of such datasets are SemEval-2010 (Kim et al., 2010) and KP20k (Meng et al., 2017), which respectively comprises scientific articles and paper abstracts, both about computer science and information technology. Detailed statistics are listed in Table 1. Only two publicly available datasets, that we are aware of, contain news documents: DUC-2001 (Wan and Xiao, 2008) and KPCrowd (Marujo et al., 2012). Originally created for the DUC evaluation campaign on text summarization (Over, 2001), the former is composed of 308 news annotated by graduate students. The latter includes 500 news annotated by crowdsourcing. Both datasets are very small and contain newswire articles from various online sources labelled by non-expert annotators, in this case readers, which is not without issues.

Thus, unlike author annotations, those produced by readers exhibit significantly lower missing keyphrases, that is, gold keyphrases that do not occur in the content of the document. In the DUC-2001 dataset for example, more than 96% of the gold keyphrases actually appear in the documents. This confirms previous observations that readers tend to assign keyphrases in an extractive fashion (Wang et al., 2015), which makes these datasets less suitable for the task at hand (keyphrase generation) but rather relevant for a purely extractive task (keyphrase extraction). Yet, author-assigned keyphrases commonly found in scientific paper datasets are not perfect either, as they are less constrained (Sood et al., 2007) and include seldom-used variants or misspellings that negatively impact performance. One can see there is an apparent lack of sizeable expert-annotated data that enables the development of neural keyphrase generation models in a domain other than scholarly texts. Here, we fill this gap and propose a large-scale dataset that includes news texts paired with manually curated gold standard annotations.

## 3 Building the *KPTimes* dataset

To create the *KPTimes* dataset, we collected over half a million newswire articles by crawling selected online news websites. We applied heuristics to identify the content (title, headline and body) of each article and regarded the keyphrases provided in the HTML metadata as the gold standard. A cherry-picked sample document is showcased in Figure 1, it allows to show present and absent keyphrases, as well as keyphrase variants (in this example `News media` and `journalism`).

We use the New York Times[1] as our primary source of data, since the content tagging policy that it applies is rigorous and well-

[1] https://www.nytimes.com/

| | Dataset | Ann. | #Train | #Dev | #Test | #words | #kp | len kp | %abs |
|---|---|---|---|---|---|---|---|---|---|
| Scholar | SemEval-2010 | $A \cup R$ | 144 | - | 100 | 7 961 | 14.7 | 2.2 | 19.7 |
| | KP20k | $A$ | 530K | 20K | 20K | 176 | 5.3 | 2.6 | 42.6 |
| News | DUC-2001 | $R$ | - | - | 308 | 847 | 8.1 | 2.0 | 3.7 |
| | KPCrowd | $R$ | 450 | - | 50 | 465 | 46.2 | 1.1 | 11.2 |
| | *KPTimes* (this work) | $E$ | 260K | 10K | 10K | 921 | 5.0 | 1.5 | 54.7 |
| | *JPTimes* (this work) | $A$ | - | - | 10K | 648 | 5.3 | 1.3 | 28.2 |

Table 1: Statistics of available datasets for keyphrase generation. Gold annotation is performed by authors ($A$), readers ($R$) or editors ($E$). The number of documents in the training (#Train), validation (dev) and testing (#Test) splits are shown. The average number of keyphrases (#kp) and words (#words) per document, the average length of keyphrases (len kp) and the ratio of keyphrases in the reference that do not appear in the document (%abs) are computed on the test set.

documented[2]. The news articles are annotated in a semi-automatic way, first the editors revise a set of tags proposed by an algorithm. They then provide additional tags which will be used by a taxonomy team to improve the algorithm.

We first retrieved the URLs of the free-to-read articles from 2006 to 2017[3], and collected the corresponding archived HTML pages using the Internet Archive[4]. Doing so allows the distribution of our dataset using a thin, URL-only list. We then extracted the HTML body content using `beautifulsoup`[5] and devised heuristics to extract the main content and title of each article while excluding extraneous HTML markup and inline ads. Gold standard keyphrases are obtained from the metadata (field types `news_keywords` and `keywords`[6]) available in the HTML page of each article. Surface form variants of gold keyphrases (e.g. "*AIDS*; *HIV*", "*Driverless Cars*; *Self-Driving Cars*" or "*Fatalities*; *Casualties*"), which are sometimes present in the metadata, are kept to be used for evaluation purposes.

We further cleansed and filtered the dataset by removing duplicates, articles without content and those with too few (less than 2) or too many (more than 10) keyphrases. This process resulted in a set of 279 923 article-keyphrase pairs. We randomly divided this dataset into training (92.8%), development (3.6%) and test (3.6%) splits.

Restricting ourselves to one source of data ensures the uniformity and consistency of annotation that is missing in the other datasets, but it may also make the trained model source-dependent and harm generalization. To monitor the model's ability to generalize, we gather a secondary source of data. We collected HTML pages from the Japan Times[7] and processed them the same way as described above. 10K more news articles were gathered as the *JPTimes* dataset.

Although in this study we concentrate only on the textual content of the news articles, it is worth noting that the HTML pages also provide additional information that can be helpful in generating keyphrases such as text style properties (e.g. bold, italic), links to related articles, or news categorization (e.g. politics, science, technology).

## 4 Data analysis

We explored the *KPTimes* dataset to better understand how it stands out from the existing ones. First, we looked at how editors tag news articles. Figure 2 illustrates the difference between the annotation behaviour of readers, authors and editors through the number of times that each unique keyphrase is used in the gold standard. We see that non-expert annotators use a larger, less controlled indexing vocabulary, in part because they lack the higher level of domain expertise that editors have. For example, we observe that frequent keyphrases in *KPTimes* are close to topic descriptors (e.g. "*Baseball*", "*Politics and Government*") while those appearing only once are very precise (e.g. "*Marley's Cafe*", "*Catherine E. Connelly*").

[2] https://lac-group.com/rules-based-tagging-metadata/

[3] https://spiderbites.nytimes.com/

[4] https://archive.org/

[5] https://www.crummy.com/software/BeautifulSoup/

[6] The change of field name correspond to the introduction of the *keywords* tag as a W3C standard.

[7] https://www.japantimes.co.jp/

Annotations in *KPTimes* are arguably more uniform and consistent, through the use of tag suggestions, which, as we will soon discuss in §5.3, makes it easier for supervised approaches to learn a good model.

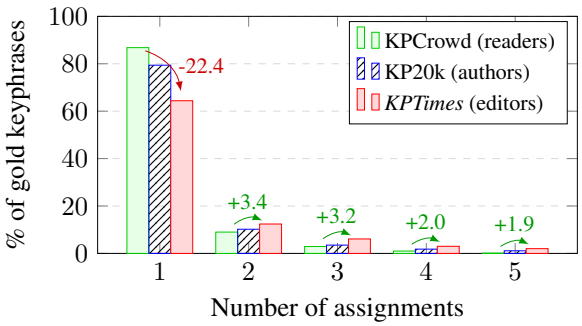


Figure 2: Distributions of gold keyphrase assignments.

Next, we further looked at the characteristics of the gold keyphrases in *KPTimes*. Table 1 shows that the number of gold keyphrases per document is similar to the one observed for KP20k while the number of missing keyphrases is higher. This indicates that editors are more likely to generalize and assign keyphrases that do not occur in the document ($\approx 55\%$). It is therefore this ability to generalize that models should mimic in order to perform well on *KPTimes*. We also note that keyphrases are on average shorter in news datasets (1.5 words) than those in scientific paper datasets (2.4 words). This may be due to the abundant use of longer, more specific phrases in scholarly documents (Jin et al., 2013).

Variants of keyphrases recovered from the metadata occur in 8% of the documents and represent 810 sets of variants in the *KPTimes* test split. These variants often refer to the same concept (e.g. "*Marijuana; Pot; Weed*"), but can sometimes be simply semantically related (e.g. "*Bridges; Tunnels*"). Thereafter, keyphrase variants will be used during model evaluation for reducing the number of mismatches associated with commonly used lexical overlap metrics.

## 5 Performance of existing models

We train and evaluate several keyphrase generation models to understand the challenges of *KPTimes* and its usefulness for training models.

### 5.1 Evaluation metrics

We follow the common practice and evaluate the performance of each model in terms of f-measure ($F_1$) at the top $N = 10$ keyphrases, and apply stemming to reduce the number of mismatches. We also report the Mean Average Precision (MAP) scores of the ranked lists of keyphrases.

### 5.2 Models

#### Baseline: FirstPhrase

Position is a strong feature for keyphrase extraction, simply because texts are usually written so that the most important ideas go first (Marcu, 1997). In news summarization for example, the lead baseline –that is, the first sentences from the document–, while incredibly simple, is still a competitive baseline (Kedzie et al., 2018). Similar to the lead baseline, we compute the **FirstPhrases** baseline that extracts the first $N$ keyphrase candidates[8] from a document.

#### Baseline, unsupervised: MultipartiteRank

The second baseline we consider, **MultipartiteRank** (Boudin, 2018), represents the state-of-the-art in unsupervised graph-based keyphrase extraction. It relies on a multipartite graph representation to enforce topical diversity while ranking keyphrase candidates. Just as FirstPhrases, this model is bound to the content of the document and cannot generate missing keyphrases. We use the implementation of MultipartiteRank available in `pke`[9] (Boudin, 2016).

#### State-of-the-art, supervised: CopyRNN

The generative neural model we include in this study is **CopyRNN** (Meng et al., 2017), an encoder-decoder model that incorporates a copying mechanism (Gu et al., 2016) in order to be able to generate phrases that rarely occur. When properly trained, this model was shown to be very effective in extracting keyphrases from scientific abstracts. CopyRNN has been further extended by (Chen et al., 2018) to include correlation constraints among keyphrases which we do not include here as it yields comparable results.

Two models were trained to bring evidence on the necessity to have datasets from multiple domains. CopySci was trained using scientific abstracts (KP20k) and CopyNews using newspaper articles (*KPTimes*), the two models use the same architecture.

[8] Sequences of adjacent nouns with one or more preceding adjectives of length up to five words.

[9] https://github.com/boudinfl/pke

| | KPCrowd | | DUC | | *KPTimes* | | *JPTimes* | | SemEval | | KP20k | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F@10 | MAP | F@10 | MAP | F@10 | MAP | F@10 | MAP | F@10 | MAP | F@10 | MAP |
| FirstPhrases | 17.1 | 16.5 | 24.6 | 22.3 | 9.2 | 8.4 | 13.5 | 13.1 | 13.8 | 10.5 | 13.5 | 12.6 |
| MultipartiteRank | 18.2 | 17.0 | 25.6 | 24.9 | 11.2 | 10.1 | 16.9 | 16.5 | 14.3 | 10.6 | 13.6 | 13.3 |
| CopySci | 15.5 | 11.1 | 12.7 | 9.7 | 11.0 | 10.6 | 18.9 | 19.8 | 20.3 | 13.8 | 25.4 | 28.7 |
| CopyNews | 8.4 | 4.2 | 10.5 | 7.2 | 39.3 | 50.9 | 24.6 | 26.5 | 7.0 | 3.5 | 6.6 | 5.1 |

Table 2: Performance on benchmark datasets composed of newspaper article, full scientific article and scientific article abstract. The generation models CopySci and CopyNews were trained respectively on KP20k and *KPTimes*. The dataset presented in this work are written in italic.

### 5.3 Results

Model performances for each dataset are reported in Table 2. Extractive baselines show the best results for KPCrowd and DUC-2001 which is not surprising given that these datasets exhibit the lowest ratio of absent keyphrases. Neural-based models obtain the greatest performance, but only for the dataset on which they were trained. We therefore see that these models do not generalize well across domains, confirming previous preliminary findings (Meng et al., 2017) and exacerbating the need for further research on this topic. Interestingly, CopyNews outperforms the other models on *JPTimes* and achieves very low scores for KPCrowd and DUC-2001, although all these datasets are from the same domain. This emphasizes the differences that exist between the reader- and editor-assigned gold standard. The score difference may be explained by the ratio of absent keyphrases that differs greatly between the reader-annotated datasets and *JPTimes* (see Table 1), and thus question the use of these rather extractive datasets for evaluating keyphrase generation.

Finally, we note that the performance of CopyNews on *KPTimes* is significantly higher than that of CopySci on KP20k, proving that a more uniform and consistent annotation makes it easier to learn a good model.

## 6 Conclusion

In this paper we presented *KPTimes*, a large-scale dataset of newswire articles to train and test deep learning models for keyphrase generation. The dataset and the code are available at `https://github.com/ygorg/KPTimes`. Large datasets have driven rapid improvement in other natural language generation tasks, such as machine translation or summarization. We hope that *KPTimes* will play this role and help the community in devising more robust and generalizable neural keyphrase generation models.

## References


James Allan. 2012. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media.

Florian Boudin. 2016. pke: an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 69–73, Osaka, Japan. The COLING 2016 Organizing Committee.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 667–672, New Orleans, Louisiana. Association for Computational Linguistics.

Jun Chen, Xiaoming Zhang, Yu Wu, Zhao Yan, and Zhoujun Li. 2018. Keyphrase generation with correlation constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4057–4066, Brussels, Belgium. Association for Computational Linguistics.

Wang Chen, Hou Pong Chan, Piji Li, Lidong Bing, and Irwin King. 2019. An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. 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 2846–2856, Minneapolis, Minnesota. Association for Computational Linguistics.

David A. Evans and Chengxiang Zhai. 1996. Noun phrase analysis in large unrestricted text for information retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Santa Cruz, California, USA. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Yiping Jin, Min-Yen Kan, Jun-Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the ACL anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790, Seattle, Washington, USA. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daum 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.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden. Association for Computational Linguistics.

Yuqing Mao and Zhiyong Lu. 2017. Mesh now: automatic mesh indexing at pubmed scale via learning to rank. *Journal of Biomedical Semantics*, 8(1):15.

Daniel Marcu. 1997. The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics*, pages 96–103.

Lus Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, and Joao P. Neto. 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592. Association for Computational Linguistics.

Paul Over. 2001. Introduction to duc-2001: an intrinsic evaluation of generic news text summarization systems.

Sanjay Sood, Sara Owsley, Kristian J. Hammond, and Larry Birnbaum. 2007. Tagassist: Automatic tag suggestion for blog posts. In *Proceedings of the First International Conference on Weblogs and Social Media, ICWSM 2007, Boulder, Colorado, USA, March 26-28, 2007*.

Daniel Tunkelang. 2009. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 855–860. AAAI Press.

Rui Wang, Wei Liu, and Chris McDonald. 2015. Using word embeddings to enhance keyword identification for scientific publications. In *Databases Theory and Applications*, pages 257–268, Cham. Springer International Publishing.

Ian H Witten, David Bainbridge, and David M Nichols. 2009. *How to build a digital library*. Morgan Kaufmann.