Keyphrase Generation: A Multi-Aspect Survey

Erion Çano, Ondřej Bojar

Charles University

Prague, Czech Republic

{cano, bojar}@ufal.mff.cuni.cz

Abstract—**Extractive keyphrase generation research has been around since the nineties, but the more advanced abstractive approach based on the encoder-decoder framework and sequence-to-sequence learning has been explored only recently. In fact, more than a dozen of abstractive methods have been proposed in the last three years, producing meaningful keyphrases and achieving state-of-the-art scores. In this survey, we examine various aspects of the extractive keyphrase generation methods and focus mostly on the more recent abstractive methods that are based on neural networks. We pay particular attention to the mechanisms that have driven the perfection of the later. A huge collection of scientific article metadata and the corresponding keyphrases is created and released for the research community. We also present various keyphrase generation and text summarization research patterns and trends of the last two decades.**

I. Introduction

A keyphrase or a keyword (here we use them interchangeably) is a short set of one or a few words that represent a concept or a topic covered in a document. They are commonly used to annotate articles or other documents and are essential for the categorization and fast retrieval of such items in digital libraries. A keyphrase string, on the other hand, is a set of comma-separated (other separators may be used as well) keyphrases associated with an article or a different type of object, describing the content and topical aspects of it.

Because of their high importance and the need to process huge amounts of documents with missing keyphrases, KG (Keyphrase Generation) attracted high academic interest since the 90s. Some basic works of that time such as [1], [2], and [3] used text features and supervised learning algorithms (popular at that time) to extract keywords from documents. Improved supervised methods like [4], [5], and [6], graph-based methods like [7], [8], and [9], or other unsupervised KG methods such as [10] and [11] were proposed in the 2000s.

Extractive KG became so popular in the 2000s and early 2010s, that the entire research field was commonly called KE (Keyphrase Extraction). This success was mainly due to the simplicity and speed of the proposed solutions. There is still a serious flaw in extractive KG: the inability to produce absent keyphrases (predicted keyphrases that do not appear in the source text). Analyzing the most popular datasets, [12] showed that present (predicted keyphrases that also appear in the source text) and absent keyphrases assigned by paper authors are almost equally frequent. Ignoring the later is thus a serious handicap.

Motivated by the advances in sequence-to-sequence applications of neural networks, several studies like [12] or [13] started to explore AKG (Abstractive Keyphrase Generation). The encoder-decoder (or sequence-to-sequence learning)

paradigm that was first utilized in the context of machine translation (e.g., in [14], [15] or [16]) got quick adaption in related tasks such as text summarization (like in [17] and [18]) or AKG. Since that time, AKG research took over and is today a vibrant field of study.

In this survey, we start by reviewing the most popular KE methods, specifically the supervised, the graph-based and the other unsupervised ones. We go on describing the popular existing keyphrase datasets and present OAGKX, a novel and huge collection of about 23 million metadata samples (titles, abstracts, and keyphrase strings) from scientific articles that is released online (http://hdl.handle.net/11234/1-3062). It can be used as a data source to train deep supervised KG methods or to create byproducts (other keyphrase datasets) from more specific scientific disciplines.

Unlike similar recent reviews such as [19], [20], or [21] that focus entirely on extractive KG (or KE), the main interest of this work in the more recent and technically advanced AKG studies which are examined in details. Particular attention is paid to the network structures and the enhancement mechanisms, as well as to the evaluation process the authors follow. We also describe certain research patterns that we observed such as the interesting analogy with similar developments in text summarization research.

II. Extractive Keyphrase Generation

Extractive keyphrase generation methods are simpler and appeared in the literature in the late 90s. They usually follow two steps. First, candidate phrases are selected from the document. Different strategies are latter applied to decide if each candidate is a keyphrase or not. The following subsections briefly describe the most popular extractive methods. More comprehensive and detailed reviews entirely focused on KE can be found in other surveys such as [19].

A. Supervised Methods

One of the first studies that considered KG as a supervised learning problem was [3]. In that study, the author experimented with texts from journal articles, email messages, and Web pages. Some of the used features were *word frequency*, *phrase length*, *number of words in phrase*, etc. C4.5 decision tree algorithm of [22] was utilized as a classifier, in combination with a bagging procedure based on random sampling with replacement presented at [23]. The author also experimented with GenEx, an algorithm described in [2] that was specifically designed to extract document keyphrases. He concluded that domain knowledge is highly valuable in the keyphrase extraction process and GenEx (using that knowledge) performs significantly better than C4.5 (not using it).

This work encouraged other researchers to develop supervised learning methods for solving KE problems.

Almost at the same time, KEA (Keyphrase Extraction Algorithm), a language-independent supervised KG algorithm was developed and presented in [1]. It uses features like *TF-IDF* and *first occurrence* and then applies Naïve Bayes classifier to determine if candidate phrases are keyphrases or not. Authors evaluate KEA using NZDL dataset (see Table I) and report that it is able to correctly identify one or two of the first five author keyphrases.

The development of Maui, a similar algorithm presented in [5] was a further step forward. Maui extends KEA in several ways. It combines more feature types and exploits Wikipedia articles as a source of linguistic knowledge. Furthermore, Maui can work well with both Naïve Bayes and bagged decision tree classifiers.

Some attempts explore various feature setups for improving the existing methods. The author in [24] investigates the role of additional features like n-grams, *noun phrases*, *POS tags*, etc. She concludes that using words or n-grams that match *POS tag* patterns increases the recall compared to the usage of n-grams only. Furthermore, according to her results, the syntactic information of the POS tags is also important for optimizing the number of keyphrases assigned to each document. The author also creates and uses Inspec, a dataset of scientific paper metadata. Other studies like [25] or [26] that followed also experimented with scientific paper texts, a practice that is common even today.

The logical structure of a scientific article is defined in [27] as the hierarchy of its logical components like title, list of authors, abstract and sections. Authors of [6] use that logical structure to build WINGNUS by limiting the number of identified candidate phrases. They further use different features like *length of phrases*, *typeface* and *position* (in title, introduction, etc.) for training the Weka implementation of Naïve Bayes (presented in [28]) to select the best candidates. Authors conclude that using the logical structure of the scientific articles yields superior performance over methods that do not consider that information.

In [29] they experimented by adding syntactic relations extracted with the dependency parser of [30]. They also tried different classifiers like Support Vector Machines of [31] and Random Forests of [32]. According to their results, the NLP-based features improve F_1 scores of all the tested methods. They also concluded that Random Forest is a good trade-off between keyphrase quality and generation speed.

There were also a few studies that applied neural network structures to perform extractive KG. In [4], for example, they used a feed-forward neural network as a classifier and paid particular attention to title headings (also subheadings) and phrase repetitions. Authors of [33], on the other hand, utilized a more complex neural network structure based on LSTMs (Long Short-Term Memory) to build an end-to-end keyphrase extraction system that eliminates the need for manual engineering of statistical features.

B. Graph-Based Methods

From the unsupervised extractive KG methods, those based on graph computations are the most numerous. In [34] they introduced TextRank, a graph-based ranking method inspired by the PageRank algorithm of [35]. They implement the idea of "voting": a vertex that represents a word or phrase (lexical unit) links to another one, casting a vote in the later. A higher number of votes to certain words or phrases suggests that they are more important. All lexical units of the source text are ranked this way. The returned keyphrases are constructed from the top N words.

Authors of [7] use the concept of the neighborhood of a given document: a set of similar documents that expands that document. They later employ PageRank on the local graph (of a single document) or the expanded graph (of the neighborhood) to rank the words and phrases. SingleRank and ExpandRank are the names of the corresponding methods they derive. The authors report that ExpandRank is significantly better than SingleRank for any size of the neighborhood.

In [36] they follow a similar approach to formulate Cite-TextRank. Authors use the documents citing the given document (citation network) to expand it and then they apply PageRank. TopicRank defined in [9] is another improvement over TextRank. It first clusters lexical units of the document according to their topic. Afterwards, it uses a graph-based ranking model to assign scores to the topic clusters. Finally, keyphrases are generated by picking one of them from each ranked cluster.

One of the fastest available KE methods is RAKE proposed in [8]. Authors first remove punctuation and stop words and then create a graph of word co-occurrences. Candidate words are scored based on the degree and frequency of each word vertice in the graph. The top-scoring ones are returned as keyphrases. Authors report that RAKE achieves higher precision and similar recall when compared with other graph-based methods like TextRank.

PositionRank is yet another graph-based KE approach recently proposed by [37]. They construct a word-level graph where they incorporate information from positions of all word occurrences. PageRank is later used to score the words and phrases. Authors show that using positions of all word occurrences works better than using the first occurrence of each word only.

C. Other Methods

Besides the two categories above, there are also other unsupervised methods that are not graph-based. They mostly utilize clustering and various similarity measures to find the best keyphrases. A very simple scheme uses TF-IDF to compute scores and rank text phrases of the entire document. This raw approach is one of the most frequent baselines in other studies that propose KG methods.

Authors of [38] proposes another basic approach based on term frequencies and stopword filtering. In [10] they argue that KG systems should be unsupervised and domain-independent. They build a KG system based on loosely structured ontologies. Authors of [39] rely on Deep Belief Networks described

in [40] to capture the intrinsic representations of documents and using them to extract keyphrases.

Another peculiar approach is the one by [41] who consider keyphrasing as a form of translation from the language of the document to the language of keyphrases. They use word alignment from statistical machine translation to learn matching probabilities between document words and keyphrase words.

Statistical language models are also used by [42] who utilize Kullback-Leibler divergence described in [43] to create a single score (including phraseness and informativeness) for ranking extracted phrases. YAKE! presented in [11] is another example of an unsupervised and feature-based extractive KG solution. They utilize features like *casing*, *word position*, *word frequency*, and more, combined in a complex scoring function that is used to yield the ranked keyphrases.

There is also a recent attempt in [44] to use the concept of word embeddings in the context of the unsupervised KE. Authors propose Key2Vec, a method for training phrase (multi-word) embeddings which are used to represent the candidate keyphrases and build the thematic representation of the document. The candidate keyphrases are later ranked based on their thematic relation with the document using the theme-weighted PageRank algorithm of [45].

The many extractive (supervised, graph-based or other) KG methods described in this section are complementary and may be used in different scenarios and for different purposes. To ease their implementation and benchmarking, the author of [46] created PKE, a Python toolkit available online (https://github.com/boudinfl/pke). It implements many of the above methods, offers pretrained and ready to use KE models and can also be easily extended to implement or benchmark new methods.

III. KEYPHRASE DATASETS

A. Popular Corpora

The recent open data initiatives and data science competitions have encouraged the creation and sharing of more and more datasets. There are papers like [47] that release data about movies, [48] about music, [49] about books and [50] that describes data of other object categories. The computational linguistics or natural language processing datasets consist of various text collections that are used to solve particular tasks. In the realm of KG, the most popular in the literature are the collections of scientific articles shown in Table I.

Inspec is one of the earliest datasets, released in [24] where the role of various linguistic features in KE is explored. It consists of 2000 paper titles (1500 for training and 500 for testing), abstracts and keywords from journals of Information Technology, published from 1998 to 2002.

One of the smallest is NUS of [51], consisting of 211 conference papers. Each paper has two sets of keyphrases: one set by the authors and a second that was created by volunteer students. Another small dataset is SemEval (or SemEval-2010) described in [52]. It is composed of 244 papers, 144 for training and 100 for testing. They were collected from ACM Digital Library and belong to conference and workshop proceedings.

TABLE I. PUBLIC KEYPHRASE DATASETS

Reference	Name	Content	# Docs
[24] Hulth	Inspec	Papers	2000
[51] Nguyen	NUS	Papers	211
[52] Kim	SemEval	Papers	244
[7] Wan	DUC	News	308
[29] Krapivin	Krapivin	Papers	2304
[53] Zhang	Twitter	Tweets	147K
[12] Meng	KP20k	Papers	567K
[1] Witten	NZDL	Reports	1800

Krapivin, the dataset released in [29] has the advantage of providing full paper texts together with the corresponding metadata. There is a total of 2304 Computer Science articles published by ACM from 2003 to 2005. The parts of each text such as title, abstract and sections are separated and marked to ease the extraction of various keyphrases.

The most popular KG dataset of the recent years is probably KP20k released in [12]. It consists of 567830 Computer Science articles, 527830 for training, 20K for validation and 20K for testing. KP20k has been used for training and evaluating various recent abstractive methods. The biggest keyphrase dataset is probably OAGK recently released in [59]. It contains 2.2M titles, abstracts and keyphrase strings of scientific papers from different disciplines.

The above scientific paper datasets are summarized in Table I. There are also a few more datasets of other document types, but they are less popular in the literature. One of them is NZDL, a collection of 1800 Computer Science technical reports, 1300 for training and 500 for testing. It is described in [1]. Authors use it to benchmark KEA, their extractive method which was one of the first.

From the news domain, the DUC (or DUC-2001) dataset of [7] is somehow popular. It consists of 308 news articles and 2048 keyphrase labels and has been used in a few extractive and abstractive KG methods. In [53] they create a dataset of about 147K tweets and their corresponding tags. Authors use it to evaluate their model for hashtag prediction. Authors of [54] use a dataset of 815 Web pages and the corresponding extracted keywords for addressing advertisements.

The two most recent datasets are probably StackExchange (post topics) and TextWorld (game observations and commands) created and used by [55]. Similar datasets can be found in other works like [56], [57] or [58].

B. A Novel and Huge Data Collection

Experimenting with keyphrases of scientific papers seems an ongoing trend that is greatly motivated by the availability of data in online academic repositories. Following the examples of [59] and [60], we took the initiative to produce an even larger collection of scientific paper keywords, titles and abstracts. Exploiting the whole data of Open Academic Graph (described in [61] and [62]), we retrieved *keywords*, *title* and *abstract* data wherever they were available. A language filter was applied to remove every text record not in English. We also lowercased and utilized Stanford CoreNLP of [63] to tokenize the *title* and *abstract* texts.

TABLE II. TOKEN STATISTICS OF OAGKX

Attribute	Title	Abstract	Keywords
Total	290 M	4 B	270 M
Min / Max	3 / 25	50 / 400	2 / 60
Mean	12.8 (4.9)	175.1 (86.5)	11.9 (7.5)
Overlaps	78 % (17 %)	68 % (25 %)	

TABLE III. KEYWORD STATISTICS OF OAGKX

Attribute	Value
Total	133 295 056
Min / Max	2 / 12
Mean	5.9 (3.1)
Present	52.7 % (28.3 %)
Absent	47.3 % (28.3 %)

Since there were several articles with very short or very long text fields (outliers), we removed any record with a title not within 3-25 tokens, abstract not within 50-400 tokens or keyphrase strings not within 2-60 tokens. We also removed records with a number of keyphrases now within 2-12. The obtained dataset is OAGKX, a collection of about 23 million article metadata records.

Some basic statistics regarding the distribution of tokens in *title*, *Abstract* and *Keywords* fields of the articles can be found in Table II. As we can see, the average lengths are about 13 tokens for the titles, 175 tokens for the abstracts, and 12 tokens for the keyphrase strings (standard deviation is given in parenthesis). We also computed the token overlaps between abstracts and titles, and between abstracts and keyphrase strings. The overlap $o(s,t) = \frac{|\{s\} \cap \{t\}|}{|\{t\}|}$ between two token vectors s (source) and t (target) is the fraction of unique tokens in t that overlap with a source token in s. As we can see, there is high repetition of abstract words, both in titles (78 %) and in keyphrases (68 %).

We further observed the distribution of keyphrases. The corresponding statistics are shown in Table III. There is a total of about 133 million keyphrases with an average of about 6 in each article. The minimal and maximal of keyphrases in each record is 2 and 12 respectively. In KG experiments, it is also important to check the frequencies of the keyphrases that are present and absent in the source texts. The present fraction $p(s,k) = \frac{|k \cap s|}{|k|}$ is the fraction of the keyphrases k that do appear in the source text s. The absent fraction $a(s,k) = \frac{|k| - |k \cap s|}{|k|}$ is the its complement, or in other words the fraction of the keyphrases k that do not appear in the source text s. As we can see, OAGKX present and absent keyphrases are almost equally frequent (52.7 % vs. 47.3 %). This is in line with the observation of [12].

Using three extractive methods described in Section II, We performed some preliminary experiments with OAGKX data. We picked YAKE!, RAKE and TopicRank which are simple and used them with their default parameters in each implementation. Given that they are unsupervised and require test data only, we picked a big test cut of 100K samples from the entire OAGKX. In addition to the preprocessing steps described above which were performed on entire OAGKX collection, we also replaced digit symbols with # and joined each title and abstract in common source string. The length of this source string was limited to 260 tokens (a paper abstract and the title should not be longer).

For the evaluation, we used F_1 score of full matches between predicted keyphrases from each method and those available in the data record (author keyphrases). We computed F_1 scores on top 5, top 7 and top 10 returned keywords. Before comparing, both sets of terms were stemmed with Porter Stemmer and duplicates were removed. The obtained

results are presented in Table IV. As we can see, the best of the three methods is YAKE, with top F_1@10 score of 21.86 %. We also observed that RAKE was considerably faster than the two other methods.

To have an idea about the topic distribution of OAGKX articles, we inspected a few randomly picked data records. We noticed that they belong to various scientific disciplines, with medicine (and its research directions) being dominant. There are also many papers about economics, social sciences or different technical disciplines. To our best knowledge, this is the biggest available collection of scientific paper data and the corresponding keyphrases. The value of OAGKX is thus twofold: (i) It can supplement the existing datasets if more training data are required. (ii) It can serve as a data source for creating scientific article subsets of more specific scientific disciplines or domains.

IV. ABSTRACTIVE KEYPHRASE GENERATION

In this section, the recent AKG methods based on the encoder-decoder framework are examined in detail. Table V summarizes some of their neural network properties, together with the evaluation data and metrics used by the authors.

A. Basic Neural Network Models

The authors of [53] were among the first to try RNNs (Recurrent Network Networks) for generating keyphrases (actually hashtags) of tweets. They adopt a joint-layer RNN with two hidden layers and two output ones. The latter are combined to form the objective layer (keyword or not). Authors build and refine a big dataset of tweets and the corresponding hashtags (keywords in this context) for evaluating their method. The basic LSTM of [64] and AKET, a tool for keyword extraction on tweets described in [65] are used as comparison baselines. Superior scores of 80.74 %, 81.19 % and 80.97 % are reported in terms of P (Precision), R (Recall) and F_1 respectively.

Another important work is [12], the first to adapt the encoder-decoder framework for AKG. Their CopyRNN model has an encoder that creates a hidden representation of the source text and a decoder that generates the keyphrases based on that representation. They employ a bidirectional GRU of [14] as the encoder and a forward GRU as the decoder. Keyphrase generation involves a beam search described in [66] with max depth 6 and beam size 200. The attention mechanism of [66] and copying mechanism of [67] are implemented to improve performance and alleviate the out-of-vocabulary words problem.

Authors evaluate CopyRNN on Inspec, Krapivin, NUS and SemEval and KP20k (IKNSK for short) datasets. Comparing with previous extractive approaches, they report state-of-the-art results in terms of F_1@5 (0.328 on KP20k) and F_1@10

TABLE IV. KE SCORES ON OAGKX (100K)

Method	F_1@5	F_1@7	F_1@10
YAKE!	19.27	21.49	21.86
RAKE	14.39	17.51	18.22
TopicRank	16.68	20.12	20.14

(0.255 on KP20k) scores for present keyphrases. They also report top scores on R@10 and R@50 for absent keyphrases. Their work created a roadmap of using the encoder-decoder framework for AKG that has been followed by many other researchers in these last three years.

In [68] they tried to optimize the speed of CopyRNN building CopyCNN made up of CNNs (Convolutional Neural Networks) which work in parallel. CNN layers are stacked on top of each other to process variable-length input text representations and gated linear units are used as the non-linearity function, same as in [69]. They also use position embeddings combined with input word embeddings to preserve the sequence order. Authors test their method using IKNSK and compare against several extractive methods and CopyRNN. They report slightly higher performance scores (in F_1@5, F_1@10, R@10, and R@50) compared to CopyRNN of [12]. Their model is also considerably faster, with generation times at least 6.2x lower.

Furthermore, authors in [13] tried to improve another aspect of CopyRNN, handling of keyword repetitions during generation. They build their model (CovRNN) utilizing a bidirectional GRU for encoding and a forward GRU for decoding. To consider the correlation of the generated target keyphrases with each other (avoiding repetitions), they implement the coverage mechanism of [70]. Same data (training on KP20k and evaluation on IKNSK) setups are used. The authors compare against extractive methods and CopyRNN. They report slightly better results compared to CopyRNN on both present (using F_1@4 and F_1@8) and absent (using R@10 and R@50) keyphrases.

B. Enhanced and Hybrid Solutions

Many works followed, improving different aspects of AKG. Authors of [71] propose a solution for handling repetition and increasing keyphrase diversity. Besides using coverage, they also implement a review mechanism that considers the source context as well as a target context (collection of hidden states) before predicting (decoding) the next keyphrase. Same as above, they implement their model (CorrRNN) with bidirectional GRU, forward GRU and beam search. They utilize the training part of KP20k and evaluate on NUS, SemEval and Krapivin datasets, comparing against several extractive methods and CopyRNN. Given that keyphrase diversity is important, besides the typical F_1 and R metrics, they also utilize α-NDCG of [72]. The authors report improvements on all reported metrics. Peak scores of 0.318 in F_1@5 and 0.278 in F_1@10 are reached on Krapivin dataset. They also assess the generalization ability of their model by training it with articles and testing it on news using DUC dataset.

All the above methods are supervised and depend on labeled training data which are not available for certain domains. In [73] they try to overcome this limitation using two approaches. In the first one, they tag unlabeled documents with

synthetic keyphrases obtained from unsupervised methods and use them for model pretraining. The pretrained model is later tuned on the labeled data. In the second one, they use multitask learning by combining the task of AKG based on labeled data with the task of title generation (a form of text summarization) on unlabeled data.

Both tasks are implemented with a bilinear LSTM as the encoder and a plain LSTM as the decoder. In the multitask learning case, the encoder is shared by the two tasks wheres the decoders are different. Authors use KP20k as a source of labeled and unlabeled data and evaluate on IKNSK. A cross-domain test with news data (DUC dataset) is also performed. Their models outrun CopyRNN on all reported metrics (F_1@5, F_1@10 and R@10) reaching a peak score of 0.308 in F_1@5 on KP20k test set.

Authors of [74] try to inject the power of extraction and retrieval into the encoder-decoder framework. A neural sequence learning model is used to compute the probability of being a keyword for each word in the source text. Those values are later used to modify the copying probability distribution of the decoder, helping the later to detect the most important words. They also use a retriever to find documents annotated similarly which provide external knowledge for the decoder and guide the generation of the keyphrases for the given document. Finally, a merging module puts together the extracted, retrieved, and generated candidates, producing the final predictions. The authors use the same data and evaluation setup as above. They report superior scores of 0.317 in F_1@5 and 0.282 in F_1@10 for present keyphrases as well as significant improvements in R@10 scores for absent keyphrases.

Furthermore, in [75], they emphasize the important role of article title which indeed can be considered as a high-level summary of the text. Their solution (TG-Net) uses a complex encoder made up of three main parts. First, a bidirectional GRU is used to separately encode the source text (abstract + title) and the title in their corresponding contextual representations. Second, a matching layer catches the relevant title information for each context word using their semantic relation. Finally, another bidirectional GRU merges the original context and the gathered title information into a final title-guided representation. The decoder is similar to the ones described above, equipped with attention and copying. The authors train with KP20k and test on IKNSK. They report important gains over CopyRNN and CopyCNN on present keyphrases, with top scores 0.372 in F_1@5 and 0.315 in F_1@10 on KP20k test set. They also report significant improvements in absent keyphrases (higher R@10 and R@50 scores).

An attempt to improve KG diversity is found in [76] where their method produces keyphrases one at a time, considering the formerly generated keyphrases. This is achieved by using multiple decoders (each of them generates only one keyphrase) that focus on different words of the source text by subtracting the attention value derived from the previous decoder. As a result, beam searches of beam size 1 are used to get the top keyphrase from each decoder and coverage is used to have diverse words in each keyphrase. The authors train their model with KP20k (the train split) and test on Inspec, Krapivin, and KP20k (the test split). They report improvements on keyphrase diversity measured using distinct-1 and distinct-2 metrics described in [77].

TABLE V.   SUMMARY OF AKG MODEL PROPERTIES. IKNSK = {INSPEC, KRAPIVIN, NUS, SEMEVAL-2010, KP20K}, NSK = {NUS, SEMEVAL-2010, KRAPIVIN}, IKK = {INSPEC, KRAPIVIN, KP20K}, GT = GENERATION TIME.

	Method				Evaluation	
Reference	Network	Att	Copy	Cov	Data	Metrics
[53] Zhang2016	joint-layer, RNN	-	-	-	Tweets	Precision, Recall, F_1
[12] Meng2017	Enc-Dec, GRU	✓	✓	-	IKNSK	F_1@5, F_1@10, R@10, R@50
[68] Zhang2017	Enc-Dec, CNN	✓	✓	-	IKNSK	F_1@5, F_1@10, R@10, R@50, GT
[13] Zhang2018	Enc-Dec, GRU	✓	✓	✓	IKNSK	F_1@4, F_1@8, R@10, R@50
[71] Chen2018a	Enc-Dec, GRU	✓	✓	✓	NSK	F_1@5, F_1@10, R@10, N@5, N@10
[73] Ye2018	Semisup, LSTM	✓	✓	-	IKNSK	F_1@5, F_1@10, R@10
[75] Chen2018b	Enc-Dec, GRU	✓	✓	-	IKNSK	F_1@5, F_1@10, R@10, R@50
[74] Chen2019	Hybrid, GRU	✓	✓	-	IKNSK	F_1@5, F_1@10, R@10
[76] Misawa2019	MultiDec, GRU	✓	✓	✓	IKK	F_1@5, F_1@10, dist1, dist2
[78] Wang2019	NTM, GRU	✓	✓	-	Blogs	F_1@1, F_1@3, F_1@5
[55] Yuan2018	catSeq, LSTM	✓	✓	-	IKNSK	F_1@5, F_1@10, F_1@M, F_1@V
[79] Chan2019	RL, GRU	✓	✓	-	IKNSK	F_1@5, F_1@M

In [78] they create another hybrid system that infuses topical information into the encoder-decoder framework. They use an NTM (Neural Topic Model) for grasping the latent topic aspects of the input text. The later go into the decoder, together with the context representation of the input obtained by the encoder. Their learning objective is modified accordingly to balance the effects of the NTM and the KG encoder-decoder. Authors conduct experiments on blog data such as Twitter, Weibo (a Chinese microblogging website) and StackExchange. They compare tag prediction of their method against various previous methods such as CopyRNN, TG-Neg, and CorrRNN, reporting considerable improvements in terms of F_1@1, F_1@3 and F_1@5 scores.

All the above works generate a fixed number of keyphrases per document. This is not optimal and realistic. In real scientific literature, different documents are paired with keyphrase sets of different lengths. To overcome this limitation and further improve the diversity of the produced keyphrases, authors of [55] propose a seq2seq generator equipped with advanced features. They first join a variable number of key terms as a single sequence and consider it as the target for sequence generation (sequence-to-concatenated-sequences or catSeq). By decoding a single of those sequences for each sample (e.g., taking top beam sequence from beam search) their model can produce variable-length keyphrase sequences for each input sample.

For a higher diversity in output sequences, they apply orthogonal regularization on the decoder hidden states, encouraging them to be distinct from each other. Authors use the same data setup as in [12] and compare against CopyRNN and TG-Net. Besides using F_1@5, F_1@10, they also propose two novel evaluation metrics: F_1@M, where M is the number of all keyphrases generated by the model for each data point, and F_1@V, where V is the number of predictions that gives the highest F_1@V score in the validation set. Considerable improvements are achieved in terms of F_1@10 (top score 0.361), F_1@M (top score 0.362) and F_1@V (top score 0.362) on KP20k test set.

C. Reinforcement Learning Perspective

Given that the above catSeq model tends to generate fewer keywords than the ground-truth, authors of [79] reformulate it from the RL (Reinforcement Learning) perspective which has also been applied recently in several text summarization works like [80], [81] or [82] and similar seq2seq applications described in [82]. The model is stimulated to generate enough keyphrases employing an adaptive reward function that is based on recall (not penalized by incorrect predictions) in undergeneration scenarios and F_1 (penalized by more incorrect predictions) in overgeneration scenarios. They use GRU instead of LSTM but keep most of the other implementation details the same as those of [55].

The authors train on KP20k and test on IKNSK. They compare the RL-implemented catSeq, CopyRNN, and TG-Net against their original versions and report improvements from the RL implementation in all cases on both F_1@5 and F_1@M with peak scores 0.321 and 0.386 respectively. The RL perspective is thus highly effective for enhancing existing AKG methods. Another contribution of their work is the novel comparison scheme they propose, with name variation sets for each ground-truth keyphrase. If a predicted keyphrase matches any name variation of a ground-truth keyphrase, it is considered as a correct prediction.

V.   KEYPHRASING RESEARCH PATTERNS

There are several patterns regarding technical and other aspects of research that show up from time to time. In this section, we briefly summarize some of such trends we identified in KG and TS (Text Summarization) research of the last two decades.

A. Experimental Patterns

All of the primary studies we consulted perform some text preprocessing steps such as tokenization and lowercasing. Most papers do not report the tokenization utility they use. A few of them like [75] and [78] report to have used Stanford CoreNLP of [63] or NLTK (www.nltk.org) for tokenizing. It is also common to find KE studies like [46], [24], and [9] that perform POS tagging and include the tags in the feature set they utilize.

A reduced vocabulary size is important to have decent AKG resutls within a reasonable computation time. For this reason, authors of many recent AKG studies like [12], [71], [68], [73], [75] and [79] replace all digit tokens with the symbol ⟨digit⟩. Stemming is also commonly used in studies

like [12], [75], [74], [24], [71] and [73] to have the predicted and golden keywords properly compared during evaluation. A stemmer that is reported is the one of [83]. There are still a few works like [13] that do not report to use stemming or any other transformation in the evaluation step.

The motivation or objective of the authors is the same in most of the studies: producing meaningful and accurate keyphrases that are similar to those set by humans which are used as ground-truth. Besides that, there are a few studies such as [76] or [71] that aim for a higher diversity or avoiding duplicates in the produced keyphrases. Producing a different number of keyphrases for each document is another requirement. It was met just recently by the model of [55].

Overcoming the need for labeled or domain-specific data was also important for certain studies like [73] and [10]. Few works such as [8] and [68] focus on computational efficiency and generation speed while trying to keep state-of-the-art accuracy. Other works such as [79] and [33] are based on neural networks and attempt to generate more keyphrases (the former) or automate feature crafting (the latter). Finally, [46] creates a framework for implementing popular methods instead of proposing a new one.

All studies do perform a formal evaluation of their contribution with the exception of [11] where they highlight the functional features of their method by means of a practical demonstration. In the evaluation phase, they usually compare with similar methods used as baselines. Regarding the choice of baselines, we observed a similar trend in both extractive and abstractive KG studies. The earlier extractive works such as [1], [6] or [26] do not compare against other methods. In few cases such as in [24] and [7], they compare different versions (or configuration choices) of their basic method.

The more recent extractive works like [1], [8], [34], [29], [33], [36], and [37] compare against the earlier ones. Similarly, the earlier abstractive KG studies such as [12] and [53] compare against extractive methods only. Instead, some of the latest abstractive works such as [75], [13] or [79] compare against both extractive and abstractive KG methods.

B. Keyphrasing vs. Summarizing

Some interesting research patterns we observed are related to the strict analogy between the dynamics of TS and KG research in the last two decades. Extensive research began in the late 90s on both tasks. Early TS works were mostly extractive, same as the KG works of the same time (commonly called KE studies). They were usually based on lexical resources and features, clustering algorithms and similarity measures (e.g., [84], [85] or [86]). Several supervised TS works such as [87] and [88] or graph-based TS works like [89], [90] and [91] bloomed, in full analogy with the KG works of Sections II-A and II-B.

The same development path has been followed in the case of abstractive studies as well. The encoder-decoder framework equipped with attention was first used by [92] for title generation. In analogy with the studies of Section IV-B, many studies like [17] or [93] added copying mechanism whereas [18] was the first that used coverage. All these innovations significantly improved the results. The trend towards the RL

approach makes no exception. It was first introduced in text summarization studies like [81] and [94]. As described in Section IV-C, It has been applied in AKG just recently.

There are still a few differences between TS and KG research that are related to the nature of these tasks. First, as presented in Section III-A, KG research works have mostly used scientific paper data. TS studies, on the other hand, have been mostly based on news articles (e.g., [95], [96] or [97]). In fact, most of the popular TS datasets like those described in [98], [99], and [93] are made up of online news articles preprocessed by the authors.

Another difference lies in the metrics that are used to perform the evaluation of the two tasks. KG methods are usually assessed by means of F_1 and recall whereas TS studies use more complex scores such as ROUGE of [100] or sometimes even BLEU of [101].

VI. Discussion

This study presents a survey of the earlier extractive KG methods and the recent cutting-edge abstractive ones that are based on the encoder-decoder framework. We first describe in brief some of the pivotal KE works which are supervised, unsupervised or graph-based. They were very successful and shaped the research field in the 2000s, mainly because of their speed and simplicity.

We then present the available keyphrase datasets that are popular in the literature and describe OAGKX, a huge article data collection that is released with this paper. It can be used as a data supplement for training deep learning models that require millions of samples. It might as well serve as a source for creating derivative datasets of scientific articles from more specific research disciplines.

The shift to the recent abstractive methods was mainly pushed from the need to annotate documents with keyphrases that do not necessarily appear in the original text. The availability of the easy-to-implement encoder-decoder framework was another motive. Advanced mechanisms such as attention, copying and coverage were added one by one and improved not only the accuracy but also the diversity of the produced keyphrases.

We further observed several similar patterns between TS and KG research. They include the transit from extractive to abstractive strategies, the use of technically advanced mechanisms (e.g., attention, copying, and coverage), and the reformulation of the methods from the reinforcement learning perspective. The latter trend is very promising and we expect to see many works in the near future exploring it in several ways for achieving different goals.

Acknowledgment

REFERENCES

[1] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, C. G. Nevill-Manning, "Kea: Practical automatic keyphrase extraction", *In Proceedings of the Fourth ACM Conference on Digital Libraries*, DL '99, pages 254–255, 1999.

[2] P. Turney, "Learning to extract keyphrases from text", unpublished.

[3] P. Turney, "Learning algorithms for keyphrase extraction", *Information Retrieval*, 2(4):303–336, May 2000.

[4] J. Wang, H. Peng, J. Hu, "Automatic keyphrases extraction from document using neural network", *Advances in Machine Learning and Cybernetics*, pages 633–641, 2006.

[5] O. Medelyan, *Human-competitive automatic topic indexing*. The University of Waikato, Phd Thesis, 2009.

[6] T. D. Nguyen, M. Luong, "Wingnus: Keyphrase extraction utilizing document logical structure", *In Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 166–169, Stroudsburg, PA, USA, 2010.

[7] X. Wan, J. Xiao, "Single document keyphrase extraction using neighborhood knowledge", *In Proceedings of the 23rd National Conference on Artificial Intelligence*, AAAI '08, Volume 2, pages 855–860, 2008.

[8] S. Rose, D. Engel, N. Cramer, W. Cowley, "Automatic keyword extraction from individual documents", *Text Mining. Applications and Theory*, pages 1–20, 2010.

[9] A. Bougouin, F. Boudin, B. Daille, "Topicrank: Graph-based topic ranking for keyphrase extraction", *In Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, 2013.

[10] D. D. Nart, C. Tasso, "A domain independent double layered approach to keyphrase generation", *In Proceedings of International Conference on Web Information Systems and Technologies*, 2014.

[11] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, A. Jatowt, Brin, "Yake! collection-independent automatic keyword extractor", *Advances in Information Retrieval*, pages 806–810, Springer International Publishing, 2018.

[12] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, Y. Chi, "Deep keyphrase generation", *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 582–592, 2017.

[13] Y. Zhang, W. Xiao, "Keyphrase generation based on deep seq2seq model", *IEEE Access*, 6:46047–46057, 2018.

[14] K. Cho, B. Merri¨nboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation", *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[15] I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks", *Advances in Neural Information Processing Systems*, 27, pages 3104–3112, 2014.

[16] D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate", *CoRR, abs/1409.0473*, 2014.

[17] S. Chopra, M. Auli, A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks", *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, 2016.

[18] A. See, P. J. Liu, C. D. Manning, "Get to the point: Summarization with pointergenerator networks", *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083, 2017.

[19] E. Papagiannopoulou, G. Tsoumakas, "A review of keyphrase extraction", *CoRR, abs/1905.05044*, 2019.

[20] K. S. Hasan, V. Ng, "Automatic keyphrase extraction: A survey of the state of the art", *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1262–1273, 2014.

[21] S. Siddiqi, A. Sharan, "Keyword and keyphrase extraction techniques: a literature review", *International Journal of Computer Applications*, 109(2), 2015.

[22] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[23] J. R. Quinlan, "Bagging, boosting, and c4.s", *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI '96, Volume 1, pages 725–730, 1996.

[24] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge", *In Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223, 2003.

[25] C. Wu, M. Marchese, J. Jiang, A. Ivanyukovich, Y. Liang, "Machine learning-based keywords extraction for scientific literature", *Journal of Universal Computer Science*, 13(10):1471–1483, October 2007.

[26] R. Bhowmik, "Keyword extraction from abstracts and titles", *In IEEE SoutheastCon 2008*, pages 610–617, April 2008.

[27] S. Mao, A. Rosenfeld, T. Kanungo, "Document structure analysis algorithms: a literature survey", *In Document Recognition and Retrieval X*, pages 197–207, Santa Clara, California, USA, January 2003.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The weka data mining software: An update", *SIGKDD Explorations Newsletter*, 11(1):10–18, November 2009.

[29] M. Krapivin, A. Autayeu, M. Marchese, E. Blanzieri, N. Segata, "Keyphrases extraction from scientific documents", *The Role of Digital Libraries in a Time of Global Change*, 2010.

[30] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kubler, S. Marinov, E. Marsi, "Maltparser: A language-independent system for data-driven dependency parsing", *Natural Language Engineering*, 13(2):95–135, 2007.

[31] C. Cortes, V. Vapnik, "Supportvector networks", *Machine Learning*, 20(3):273–297, September 1995.

[32] L. Breiman, "Random forests", *Machine Learning*, 45(1):5–32, October 2001.

[33] J. Villmow, M. Wrzalik, D. Krechel, "Automatic keyphrase extraction using recurrent neural networks", *Machine Learning and Data Mining in Pattern Recognition*, pages 210–217, 2018.

[34] R. Mihalcea, P. Tarau, "TextRank: Bringing order into texts", *In Proceedings of EMNLP-04 and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[35] S. Brin, L. Page, "The anatomy of a large-scale hypertextual web search engine", *Comput. Netw. ISDN Syst.*, 30(1–7):107–117, April 1998.

[36] S. D. Gollapalli, C. Caragea, "Extracting keyphrases from research papers using citation networks", *In Proceedings of the TwentyEighth AAAI Conference on Artificial Intelligence*, AAAI '14, pages 1629–1635, 2014.

[37] C. Florescu, C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents", *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1105–1115, 2017.

[38] Y. HaCohen-Kerner, "Automatic extraction of keywords from abstracts", *Knowledge-Based Intelligent Information and Engineering Systems*, pages 843–849, 2003.

[39] T. Jo, J. Lee, "Latent keyphrase extraction using deep belief networks", *International Journal of Fuzzy Logic and Intelligent Systems*, 15(3):153–158, 2015.

[40] G. Hinton, S. Osindero, Y. Teh, "A fast learning algorithm for deep belief nets", *Neural Computation*, 18(7):1527–1554, July 2006.

[41] Z. Liu, X. Chen, Y. Zheng, M. Sun, "Automatic keyphrase extraction by bridging vocabulary gap", *In Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 135–144, Stroudsburg, PA, USA, 2011.

[42] T. Tomokiyo, M. Hurst, "A language model approach to keyphrase extraction", *In Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, MWE '03, Stroudsburg, PA, USA, 2003.

[43] M. Vidyasagar, "Kullback-leibler divergence rate between probability distributions on sets of different cardinalities", *In 49th IEEE Conference on Decision and Control*, pages 948–953, December 2010.

[44] D. Mahata, J. Kuriakose, R. R. Shah, R. Zimmermann, "Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings", *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2, pages 634–639, New Orleans, Louisiana, USA, June 2018.

[45] A. N. Langville, C. D. Meyer, "Deeper Inside PageRank", *Internet Mathematics*, 1:3, 335-380, January 2004.

[46] F. Boudin, "pke: an open source pythonbased keyphrase extraction toolkit", *In Proceedings of COLING 2016, the 26th International Confer-*

ence on Computational Linguistics: System Demonstrations, pages 69–73, Osaka, Japan, December 2016.

[47] F. M. Harper, J. A. Konstan, "The movielens datasets: History and context", *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, December 2015.

[48] E. Çano, M. Morisio, "Music mood dataset creation based on last.fm tags", *In Computer Science & Information Technology (CS & IT)*, pages 15–26, Vienna, Austria, May 2017.

[49] Z. Zajac, "Goodbooks-10k: a new dataset for book recommendations", *FastML*, 2017.

[50] E. Çano, M. Morisio, "Characterization of public datasets for recommender systems", *In 2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, pages 249–257, September 2015.

[51] T. D. Nguyen, M. Kan, "Keyphrase extraction in scientific publications", *In Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers*, ICADL '07, pages 317–326, 2007.

[52] S. N. Kim, O. Medelyan, M. Kan, T. Baldwin, "Automatic keyphrase extraction from scientific articles", *In Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden, July 2010.

[53] Q. Zhang, Y. Wang, Y. Gong, X. Huang, "Keyphrase extraction using deep recurrent neural networks on twitter", *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 836–845, 2016.

[54] W. Yih, J. Goodman, V. R. Carvalho, "Finding advertising keywords on web pages", *In Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 213–222, 2006.

[55] X. Yuan, T. Wang, R. Meng, K. Thaker, D. He, A. Trischler, "Generating diverse numbers of diverse keyphrases", *CoRR, abs/1810.05241*, 2018.

[56] M. Dredze, H. M. Wallach, D. Puller, F. Pereira, "Generating summary keywords for emails using topics", *In Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, pages 199–206, New York, USA, 2008.

[57] M. Grineva, M. Grinev, D. Lizorkin, "Extracting key terms from noisy and multitheme documents", *In Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 661–670, New York, USA, 2009.

[58] K. M. Hammouda, D. N. Matute, M. S. Kamel, "Corephrase: Keyphrase extraction for document clustering", *Machine Learning and Data Mining in Pattern Recognition*, pages 265–274, 2005.

[59] E. Çano, O. Bojar, "Keyphrase generation: A text summarization struggle", *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, pages 666–672, Minneapolis, Minnesota, USA, June 2019.

[60] E. Çano, O. Bojar, "Efficiency Metrics for Data-Driven Models: A Text Summarization Case Study", *CoRR, abs/1909.06618*, 2019.

[61] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, "Arnetminer: Extraction and mining of academic social networks", *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, USA, 2008.

[62] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B. Hsu, K. Wang, "An overview of microsoft academic service (mas) and applications", *In Proceedings of the 24th International Conference on World Wide Web*, WWW '2015, pages 243–246, New York, USA, 2015.

[63] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, "The Stanford CoreNLP natural language processing toolkit", *In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA, June, 2014.

[64] S. Hochreiter, J. Schmidhuber, "Long short-term memory", *Neural Computation*, 9(8):1735– 1780, November 1997.

[65] L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. M. Matos, J. Neto, J. Carbonell, "Automatic keyword extraction on twitter", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Volume 2, pages 637–643, Beijing, China, July 2015.

[66] D. Dahlmeier, H. T. Ng, 'A beam-search decoder for grammatical error correction", *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP–CoNLL '12, pages 568–578, Stroudsburg, PA, USA, 2012.

[67] J. Gu, Z. Lu, H. Li, V. O. K. Li, David, C. Christopher, V. Vapnik, "Incorporating copying mechanism in sequence-to-sequence learning", *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Volume 1, pages 1631–1640, Berlin, Germany, August 2016.

[68] Y. Zhang, Y. Fang, X. Weidong, "Deep keyphrase generation with a convolutional sequence to sequence model", *In 2017 4th International Conference on Systems and Informatics*, pages 1477–1485, November 2017.

[69] Y. N. Dauphin, A. Fan, M. Auli, D. Grangier, "Language modeling with gated convolutional networks", *In Proceedings of the 34th International Conference on Machine Learning*, ICML '17, Volume 70, pages 933–941, 2017.

[70] Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li, "Modeling coverage for neural machine translation", *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 76–85, 2016.

[71] J. Chen, X. Zhang, Y. Wu, Z. Yan, Z. Li, "Keyphrase generation with correlation constraints", *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4057–4066, Brussels, Belgium, October-November 2018.

[72] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon, "Novelty and diversity in information retrieval evaluation", *In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'08, pages 659–666, New york, USA, 2008.

[73] H. Ye, L. Wang, "Semi-supervised learning for neural keyphrase generation", *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4142–4153, Brussels, Belgium, October-November 2018.

[74] W. Chen, H. P. Chan, P. Li, L. Bing, I. King, "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, pages 2846–2856, Minneapolis, Minnesota, USA, June 2018.

[75] W. Chen, Y. Gao, J. Zhang, I. King, M. Lye, "Title-guided encoding for keyphrase generation", *CoRR, abs/1808.08575*, 2018.

[76] S. Misawa, Y. Miura, M. Taniguchi, T. Ohkuma, "Multiple keyphrase generation model with diversity", *Advances in Information Retrieval*, pages 869–876, 2019.

[77] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, "A diversity-promoting objective function for neural conversation models", *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, USA, June 2016.

[78] Y. Wang, J. Li, H. P. Chan, I. King, M. R. Lyu, S. Shi, "Topicaware neural keyphrase generation for social media language.", *CoRR abs/1906.03889*, 2019.

[79] H. P. Chan, W. Chen, L. Wang, I. King, "Neural keyphrase generation via reinforcement learning with adaptive rewards", *CoRR, abs/1906.04106*, June 2019.

[80] Y. Chen, M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting", *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Volume 1, pages 675–686, 2018.

[81] R. Paulus, C. Xiong, R. Socher, "A deep reinforced model for abstractive summarization", *CoRR, abs/1705.04304*, 2017.

[82] Y. Keneshloo, T. Shi, N. Ramakrishnan, C. K. Reddy, "Deep reinforcement learning for sequence to sequence models", *CoRR, abs/1805.09461*, 2018.

[83] M. Porter, "An algorithm for suffix stripping", *Program*, 40(3):211–218, 2006.

[84] R. Barzilay, M. Elhadad, "Using lexical chains for text summarization", *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, 1997.

[85] S. Azzam, K. Humphreys, R. Gaizauskas, "Using coreference chains fortext summarization", *in Proceedings of the Workshop on Coreference and Its Applications, CorefApp '99*, pages 77–84, Stroudsburg, PA, USA. Association for Computational Linguistics.

[86] J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, "Multi-document summarization by sentence extraction", *In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, NAACL-ANLP-AutoSum '00, Volume 4, pages 40–48, Stroudsburg, PA, USA, 2000.

[87] J. Fukumoto, "Multi-document summarization using document set type classification", *In Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, NTCIR-4, Tokyo, Japan, June 2004.

[88] K. Wong, M. Wu, W. Li, "Extractive summarization using supervised and semisupervised learning", *In Proceedings of the 22Nd International Conference on Computational Linguistics*, COLING '08, Volume 1, pages 985–992, Stroudsburg, PA, USA, 2008.

[89] X. Wan, J. Yang, "Improved affinity graph based multi-document summarization", *In Proceedings of the Human Language Technology Conference of the NAACL* NAACL '06, pages 181–184, Stroudsburg, PA, USA, 2006.

[90] I. Mani, E. Bloedorn, "Multidocument summarization by graph search and matching", *In Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI '97/IAAI '97, 622–628, 1997.

[91] G. Erkan, D. R. Radev, , C. Christopher, V. Vapnik, "Lexrank: Graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligence Research*, 22(1):457–479 December 2004.

[92] A. M. Rush. S. Chopra, J. Weston, "A neural attention model for abstractive sentence summarization", *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, 2015.

[93] R. Nallapati, B. Zhou, C. Santos, C. Gulcehre, B. Xiang, "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, 2016.

[94] Y. Chen, M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting", *CoRR, abs/1805.11080*, May 2018.

[95] K. McKeown, D. R. Radev, "Generating summaries of multiple news articles", *In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 74–82, New York, USA, 1995.

[96] K. Kaikhah, "Automatic text summarization with neural networks", *2nd International IEEE Conference on Intelligent Systems*, Volume 1, pages 40–44, 2004.

[97] S. Harabagiu, F. Lacatusu, "Topic themes for multi-document summarization", *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 202–209, New York, USA, 2005.

[98] M. Grusky, M. Naaman, Y. Artzi, , "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*, pages 708–719, 2018.

[99] G. David, C. Christopher, V. Vapnik, "English gigaword", *Linguistic Data Consorium*, Philadelphia, USA, 2003.

[100] C. Lin, "Rouge: A package for automatic evaluation of summaries", *In Proceedings of ACL workshop on Text Summarization Branches Out*, page 10, 2004.

[101] K. Papieni, S. Roukos, T. Ward, W. J. Zhu, "Bleu: A method for automatic evaluation of machine translation", *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002.