Information Processing and Management 56 (2019) 102088

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Textual keyword extraction and summarization: State-of-the-art

Zara Nasar, Syed Waqar Jaffry, Muhammad Kamran Malik

Artifical Intelligence and Multi-Disciplinary Research Lab, National Center of Artificial Intelligence, Punjab University College of Information Technology, University of the Punjab, 54000, Lahore, Pakistan

ARTICLE INFO

Keywords:
Automatic keyword extraction
Text summarization
Deep Learning

ABSTRACT

With the advent of Web 2.0, there exist many online platforms that results in massive textual data production such as social networks, online blogs, magazines etc. This textual data carries information that can be used for betterment of humanity. Hence, there is a dire need to extract potential information out of it. This study aims to present an overview of approaches that can be applied to extract and later present these valuable information nuggets residing within text in brief, clear and concise way. In this regard, two major tasks of automatic keyword extraction and text summarization are being reviewed. To compile the literature, scientific articles were collected using major digital computing research repositories. In the light of acquired literature, survey study covers early approaches up to all the way till recent advancements using machine learning solutions. Survey findings conclude that annotated benchmark datasets for various textual data-generators such as twitter and social forms are not available. This scarcity of dataset has resulted into relatively less progress in many domains. Also, applications of deep learning techniques for the task of automatic keyword extraction are relatively unaddressed. Hence, impact of various deep architectures stands as an open research direction. For text summarization task, deep learning techniques are applied after advent of word vectors, and are currently governing state-of-the-art for abstractive summarization. Currently, one of the major challenges in these tasks is semantic aware evaluation of generated results.

1. Introduction

Due to advent of Word Wide Web (WWW) and later Web 2.0, there currently exists wide variety of platforms that are resulting in enormous data generation. Social networking websites such as Facebook, Twitter are generating terabytes of data. Similarly question-answering websites such as Quora and StackOverFlow also data is being produced abundantly by means of social networking sites, question-answering engines and various sharing portals. It is expected that, by 2020, total data generated would be around forty four zeta-bytes (Waterford Technologies, 2017; Marr, 2019). As humans tend to communicate by means of various data forms including images, videos, sound and textual streams on various sites over the internet, this data carries a huge value. Amongst the various data types, this study is primarily focused on textual data. Many question-answering systems, news-wire agencies, blogging websites, research engines, digital libraries and e-commerce websites share most of their data in form of text. The potential information hidden in these bulks of textual data can be used in order to perform variety of tasks. For example, in the domain of e-commerce, retailers and product manufacturers can get a better idea about customer's biases by means of customer review's analysis. This type of analysis can help in optimization of products and its respective features. Similarly, in question-answering search engines, textual analysis can help in identifying keywords and generating short summaries. In the same way, by processing data in scientific articles, one can gain insights about state-of-the art research problems, techniques, available solutions, current results and open-problems. Many search

E-mail addresses: zara.nasar@pucit.edu.pk (Z. Nasar), swjaffry@pucit.edu.pk (S.W. Jaffry), kamran.malik@pucit.edu.pk (M.K. Malik).

https://doi.org/10.1016/j.ipm.2019.102088
Received 2 January 2019; Received in revised form 5 June 2019; Accepted 19 July 2019

engines tend to present a brief summary of relevant sites as per the user query. Additionally, there are many solutions available to generate document summaries. Such concise overview of documents and/or websites assist users to have brief idea about the content beforehand; consequently, saving the precious time of users. Thus, in light of above points, textual information processing can yield powerful insights from data and it also helps in saving user's time and provide potential information that could be used for betterment of humanity.

Next pertinent question is how this bulk of data can be processed. There are two major approaches to extract key insights or to present the data insights in a concise fashion. First approach is the manual analysis i.e. to manually process the data and start logging the information that seems useful. Other approach is automated analysis that would employ computing technologies in order to perform text analysis. Later approach is more appealing due to huge data sizes as processing huge amount of text in a manual fashion won't be feasible. In order to harness huge amount of data, a whole domain named Text Mining is dedicated. Text Mining offers wide variety of research problems with each having a specific goal. In the course of this particular study, two major Text Mining problems are being explored. These involve extraction of key information and presentation of key information in a brief and concise form, with former being known as automatic keyword/keyphrase extraction (AKE) and later regarded as textual summarization (TS).

Both AKE and TS are focused on presenting key-information. AKE is being widely used to annotate the articles to reflect its potential categories. It carries significance importance in domain of Information Retrieval. This is because keywords from any piece of data points suitability and relevance of respective data source to other relevant major domains and/or topics. Apart from that, AKE is also used for summarization. TS, on the other hand, deals with presenting a concise summary for a given document. TS is primarily useful to get the overview and to present the bird's eye view of a document. This information helps in saving time of users by providing it with key information within a document in a human consumable form.

The primary reason to review these two separate research areas into one is the prime similarity in their nature. As AKE deals with extraction of key information, whereas TS deals with presentation of key information in a form of a passage. Therefore, in literature, there exist many studies where AKE is performed prior to TS in order to acquire keywords (Fang, Mu, Deng, & Wu, 2017; Gayo-Avello, Alvarez-Gutierrez, & Gayo-Avello, 2004; Lynn, Choi, & Kim, 2018; Thomas, Bharti, & Babu, 2016). That motivates us to focus on these separate research problems into one comprehensive study. In addition, each of the problems is also addressed individually in literature.

In the context of this study, both AKE and TS are extensively reviewed and the corresponding literature is further classified based on approaches. The prime focus of this study is to present overall advancements ever-since the advent of these two research problems. Though, there exist reviews in literature for each of these research problems, none of them covers recent advancements that include deep learning approaches. Therefore, neural network based approaches are being majorly focused.; In order to highlight these approaches, neural networks and their derivatives are being treated separately even though they can be classified as supervised or unsupervised approaches depending on the application.

Rest of this paper is structured as follows. Background in Section 2 covers major algorithms and datasets that are being employed to solve and evaluate the respective tasks to provide brief overview of domain. Section 3 represents the methodology opted to perform this literature survey. Sections 4 and 5 briefly explain the state-of-the-art in domains of AKE and TS respectively. Section 6 presents information regarding open-source toolkits available to perform IE and TS related tasks. Finally, Section 7 provides conclusion of this study along with future directions.

2. Background

This section is dedicated to briefly introduce both tasks i.e. AKE and TS and express the evaluation measures that are used to determine the effectiveness of any proposed approach. Another essential ingredient required for evaluation of a proposed approach is dataset. Hence, this section also covers major benchmark datasets that are being used in literature for the evaluation of AKE and TS respectively.

Keyword extraction (KE) refers to the extraction of words that best represent a document. Its purpose is to provide a brief idea that what a particular document is about. The process of automatically extracting keywords by means of computing mechanism is called automatic keyword extraction (AKE). If a user is provided with such information, it could provide a brief idea about the content presented in relevant document. Following example shows keywords, highlighted with bold italics, where respective piece of text as well as extracted keywords are taken from Sharan (2019).

"***Kendrick Lamar***, Hip-hop's Newest ***Old-School*** Star
By Dave Turner
Last updated: February 3rd, 2015

> Everybody just wants to have fun, be with the scene,' ***Kendrick Lamar*** said when we met in his cramped quarters inside the ***Barclays Center*** in ***Brooklyn*** last fall. 'Certain people get ***backstage***, people that you would never expect…You ain't with the media! You ain't into music! You ain't into sports! You're just here.' The rapper, now 27, had just finished his set as the opening act on this stretch of ***Kanye*** West's Yeezus tour, and he was sitting low in an armchair in his ***trademark*** black hoodie ***surrounded*** by exactly those people. 'Hey man, thank you again, appreciate the access back' "…

Above example highlights the importance of AKE in practical applications. It is of immense importance as it tends to provide a bird's eye view of data at hand. Thus, AKE is being widely addressed in literature, where keywords can be single word or multi-word (in the form of phrases).

Textual Summarization (TS), on the other hand, refers to process of generating summary that involves identification of key concepts residing in a text followed by the expression of these key concepts in a brief, clear and concise fashion. It concerns selection of text nuggets that provide overview of information residing in a document. In addition to that, generated summary, being concise and brief, must also preserve the information content and overall meaning of the document. Thus, TS is of great importance as information is presented in a brief fashion and users can read the crux of document in quite a short amount of time. In the following passage, text in bold italics represents those sentences that are included in final summary.

> "Thomas A. Anderson is a man living two lives. ***By day he is an average computer programmer and by night a hacker known as Neo.*** Neo has always questioned his reality, but the truth is far beyond his imagination. ***Neo finds himself targeted by the police when he is contacted by Morpheus, a legendary computer hacker branded a terrorist by the government.*** Morpheus awakens Neo to the real world, a ravaged wasteland where most of humanity have been captured by a race of machines that live off of the humans' body heat and electrochemical energy and who imprison their minds within an artificial reality known as the Matrix. ***As a rebel against the machines, Neo must return to the Matrix and confront the agents: super-powerful computer programs devoted to snuffing out Neo and the entire human rebellion.***"

2.1. Evaluation

Evaluation for AKE is usually performed by the comparison between the human annotated dataset (also known as gold standard) and system-generated results. For such comparison, evaluation metrics inspired from Information Retrieval domain are majorly used that include precision, recall and f-score. Precision presents the ratio between correctly identified entities and total amount of guessed entities by the systems. On the other hand, recall, presents the ratio between correctly identified entities and total number of entities that actually exist in gold standard dataset. In other words, precision reflects the correctness of developed system, whereas recall presents the coverage provided by the system. Table 1 shows a confusion matrix for binary classification problem, that is a widely used notion to define and present precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

where FP and FN are regarded as type-1 and type-2 error respectively. Moreover, it is usually desired to have high precision and high recall. Hence, another metric that considers both measures named f-score, is widely used. F-score can be calculated by taking the harmonic mean of recall and precision as formulated below.

$$Fscore = (1 + \beta^2)\frac{Precision*Recall}{\beta^2*Precision + Recall} \tag{4}$$

Eq. (4) enables researchers to adjust the weights of both, individual recall and precision metrics depending on the problem at hand. However, it is often desired to give both of these metrics a similar weight. This can be achieved using $\beta = 1$, and resultant metric is often regarded as balanced f-score, f1-score or f-measure. In the context of this study, f-measure is being used in order to evaluate IE constructs along with precision and recall, unless explicitly mentioned. Average matching on the other hand, computes number of matches on predicted set and gold-standard set, and later computes average over total test data. For AKE evaluation, metrics are reported with various number of machine generated keywords. Widely used numbers for evaluation include five, ten, fifteen and twenty keywords.

Evaluation of generated summaries is often carried out via ROUGE (Lin, 2004; Tseng, 1998) metric, which tends to measure the overlap between reference summary, that is usually generated by humans, and system generated summary. There are three widely used ROUGE variants ROUGE-1 and ROUGE-2 where results are computed with respect to unigrams, bigrams and longest common-subsequence respectively. Generic formulation for ROUGE-N is expressed in Eq. (5). Another widely used metrics is ROUGE-L, where L denotes longest common subsequence (LCS). Eqs. (6)–(8) presents formulation for ROUGE-L when summary X of length m is compared with summary Y having length n.

$$ROUGE - N = \frac{\sum_{S \in Reference\ Summaries} \sum_{gram_k \in S} Count_{match}(gram_k)}{\sum_{S \in Reference\ Summaries} \sum_{gram_k \in S} Count(gram_k)} \tag{5}$$

Table 1
A confusion matrix for binary classification.

	Predicted (True)	Predicted (False)
Actual (True)	True Positive (TP)	False Negative (FN)
Actual (False)	False Positive (FP)	True Negative (TN)

$$P_{lcs} = \frac{LCS(X, Y)}{m} \tag{6}$$

$$R_{lcs} = \frac{LCS(X, Y)}{n} \tag{7}$$

$$ROUGE - L = (1 + \beta^2)\frac{P_{lcs} * R_{lcs}}{\beta^2 * P_{lcs} + R_{lcs}} \tag{8}$$

Reference summaries in Eq. (5) refer to gold standard dataset of summaries. $Count_{match}(gram_k)$ refers to maximum number of n-grams that are present in both i.e. reference summary as well as candidate summary. On the other hand, $Count(gram_k)$ represents the total number of n-grams in any candidate summary. Overall, ROUGE metric is more related with the concept of recall.

In conclusion, widely used evaluation metrics for AKE include precision, recall and f-score; whereas ROUGE measure and its variants are being widely used to evaluate text summarization. All the results against evaluation measures that include precision, recall, f-score, accuracy and ROUGE are reported in terms of percentages. P/R/F is used in rest of the paper to denote precision, recall and f-score scores reported in various studies.

2.2. Datasets

Datasets serve as the primary ingredient in text mining tasks as they are used for training and testing of various techniques. This section describes major benchmark datasets that are being widely used in literature for training and eventual evaluation of various proposed techniques to perform AKE and TS.

2.2.1. Datasets for AKE

There exist multiple datasets for AKE. Every dataset is based on a specific domain or genre. Many datasets are focused on performing AKE on news articles as well as scientific articles. Whereas, there also exist datasets that are focused on AKE from text resources that are short in nature such as emails and tweets. Following section compiles widely used datasets for AKE.

2.2.1.1. SemEval. Semantic Evaluation is a yearly workshop that is focused towards solution of semantic-oriented problems. Its repository comprises various datasets that are being widely used to perform different text mining tasks. Major SemEval tasks that are relevant to keyword extraction include SemEval 2010. SemEval 2010 includes one task of keyword extraction from scientific articles, where articles included in collection are conference and workshop papers from ACM library. Each paper spans from six to eight pages. To ensure diversity in the corpus, articles belonged to various domains including distributed systems, multiagent systems, information retrieval and economics.

2.2.1.2. Inspec. Another major scientific article dataset is presented in Hulth (2003) known as Inspec dataset that consists of 2000 English abstracts acquired from Inspec database. Included abstracts cover studies from 1998 to 2002 and include journal articles belonging to the domains of computer science and information technology. It is one of the highly used dataset for AKE evaluation.

2.2.1.3. ACM[1] dataset. There exist multiple datasets that use Association for Computing Machinery (ACM) data. One such dataset comprising around 2304 Computer Science Journal articles is presented in Krapivin and Marchese (2009). The acquired papers lie in the range of 2003 to 2005. Furthermore, all acquired articles are stored in text format following UTF-8 encoding. Each article consists of title, abstract, body, references and citations. Kumar and Srinathan (2008) also uses 140 documents from ACM with query “human factors” to collect full-length articles. Zhang et al. (2017) presents dataset comprising conference articles from two top-tier machine learning conferences of ACM that include Knowledge Discovery and Data Mining (KDD) and World Wide Web (WWW) for AKE.

2.2.1.4. Turney (2000). A dataset containing data from various genres including scientific papers, web pages and emails is presented in Turney (2000). Scientific articles in this dataset belong to various domains including chemistry along with brain, behavioral and neuro sciences. On the other hand, web pages are acquired from three different sources that include Aliweb, NASA and FIPS.

2.2.1.5. Wang, Peng, and Hu (2006). Two sources that include Baidu, a popular search engine available in China hosted at www.baidu.com, and google search engine; are used to search and download scientific articles in English and Chinese languages. Total of 300 articles were downloaded whereas subjects of those articles include wide variety such as economics, manage science, biology, information science, engineering science, chemistry, physics, mathematics.

2.2.1.6. Sarkar, Nasipuri, and Ghose (2012). This dataset comprises scientific articles belonging to various genres that include medical, legal and economics. Total of 210 full journal articles with pages ranging of 6–30 were collected. Some of the Medical articles’ sources include various issues of medical journals focused on topics of medicine, pediatrics, psychology, counseling, surgery, cardiology and anxiety disorders. Economics and legal articles were downloaded from various Springer and Elsevier journals. In

[1] http://dl.acm.org.

addition to these two, economics articles were downloaded from Economics Letters and The Journal of Policy Modeling; whereas The AGORA International Journal of Juridical Sciences was used for legal cases.

2.2.1.7. CogPrints[2]. CogPrints, is an electronic archive having wide coverage of papers that are pertinent in the study of cognition. Archive includes articles belonging to the domains of Psychology, Neuroscience, and Linguistics. In addition, it further carries relevant studies covering from Computer Science (e.g., neural networks, learning, speech, vison, robotics and artificial intelligence), Philosophy (e.g., logic, science, knowledge, language, mind), Medicine (e.g., Psychiatry, Neurology,), Biology (e.g., evolutionary theory, behavioral ecology), Anthropology (e.g., cognitive ethnology), as well as any other portions of the mathematical, social and physical sciences related to cognition.

Some datasets for AKE use CogPrints as source. One of these include Kumar and Srinathan (2008), which offers two datasets: one that consists of full-length text and other that is based solely on abstracts. Full-length collection carries wide variety of various research domains that include biology, animal behavior, psychology, philosophy, biophysics and AI. All Details of abstract dataset are shared in table.

2.2.1.8. Other scientific articles' based datasets. Three sets of data reported in Aquino, Hasperué, and Lanzarini (2014) employ articles from Workshop of Researchers in Computer Science (WICC) and from Argentine Congress of Computer Science (CACIC). Two datasets from WICC carry 43 and 166 documents respectively whereas third dataset from CACIC consists of 72 articles. Another dataset based on scientific articles is presented in Witten, Paynter, Frank, Gutwin, and Nevill-Manning (1999) that employs data collection from New-Zealand digital library (NZDL) containing Computer Science Technical Reports (CSTR). Total of 1800 reports were used where author has already listed the keyphrases against respective report. Another dataset aimed to train deep learning frameworks is presented in Meng et al. (2017) which consists of 567,830 articles from variety of sources including, ScienceDirect, Wiley, ACM Digital Library and Web of Science. Subset of this data is used to generate another dataset named KP20k, which carries titles, keywords and abstracts against 20,000 computer science articles.

2.2.1.9. MPQA. The MPQA Corpus (Rose, Engel, Cramer, & Cowley, 2010) is based on the data provided by the Center for the Extraction and Summarization of Events and Opinions in Text (CERATOPS) that carries total of 535 articles. This dataset contains news reported in 187 various US and foreign news sources. The news aritcles included are dated from June 2001 to May 2002.

2.2.1.10. DUC 2001 for AKE. The dataset was originally used for document summarization. It consists of 308 distinct news articles collected from TREC-9. The TREC-9 corpus comprises news data from following sources:

1) Foreign Broadcast Information Service (FBIS)
2) Los Angeles Times
3) Financial Times
4) San Jose Mercury News
5) Wall Street Journal
6) Associated Press newswire

The selected 308 articles are classified into 30 news topics and. the data was manually tagged and it carries at most 10 keyphrases per document. The annotated dataset caries total of 2488 key phrases with 8.08 and 2.09 being the average keywords/doc and average words/phrase was 2.09.

2.2.1.11. Sterckx, Demeester, Deleu, and Develder (2018). Three Belgian media companies' data in Dutch language is used to construct total of three datasets, each having a different focus. The first one is public-service broadcaster VRT. The data from VRT is subset subsets of its official news channel website "De Redactie" and its specialized sports section Sporza. Both these are used to build online news and sports dataset. The second media company, Sanoma, owns publishing of lifestyle, fashion, and health magazines, which is used to build Lifestyle Magazines test collection. Lastly, Belga, the third media company, has a digital press that is used to construct News articles dataset. Each of this collection carries 1200–2000 documents, annotated by six annotators on average.

2.2.1.12. 20 Newsgroups. (20News) dataset is a vast collection of newsgroup documents carrying total of 18,845 news articles. This huge news corpus is further classified in 20 different newsgroups. Here, each newsgroup represents a separate topic. Wide variety of topics that include space, atheism, Christianity, guns and mideast are covered in this dataset.

2.2.1.13. Obama–McCain debate (OMD). OMD dataset is based on the first United States presidential TV debate, held September 2008. Tweets generated during this debate were crawled by employing Twitter search API. Total of 3269 relevant tweets were filtered using relevant hashtags including #current, #debate08 and #tweetdebate. Major topics covered in this dataset include terrorist threat, lesson of Iraq, financial recovery, tax cut and security.

[2] http://cogprints.org.

2.2.1.14. Jo (2003). This study collected news articles to calculate inverse document frequency (IDF) against each distinct word. Data covers wide variety of fields including Wireless communication, Sports, Public, Politics, Pharmacy, Migration, Healthcare Regulation, Business and WeirdNuz,. Moreover, hundred documents were acquired against each field making total document count 900.

2.2.1.15. Summary. Some of the notable datasets in terms of size, diversity and data sources are briefly described above. There exist various other datasets in addition to aforementioned ones, which are not described for the sake of brevity. Statistics against various datasets that are employed in the reviewed literature to perform AKE are presented in Table 2. The datasets which does not carry information against more than three column headers are not included here. In following table, *Study* column presents the research article in which dataset was proposed or the dataset name. Furthermore, several datasets that are used by another study contain "Used in" clause followed by the respective study, which employed the dataset. *Genre* represents the major textual data category which was used to construct a dataset such as scientific articles, news, tweets etc. *Count* represents the total amount of items included in a dataset. *Topic* describes major areas to which selected items belong. *Source* represents various source(s) from where a particular dataset was collected. Lastly, *Language* column describes the language of respective dataset.

Following abbreviations are used in the table: Association for Computing Machinery (ACM), Workshop of Researchers in Computer Science (W ICC), Argentine Congress of Computer Science (CACIC), Center for the Extraction and Summarization of Events and Opinions in Text (CERATOPS), New-Zealand digital library (NZDL), United Nations (UN), Microsoft Academic Search (MAS), National Natural Science Foundation of China (NNSFC), Distributed Systems (DS), Information Retrieval (IR), Artificial Intelligence (AI) and Human Computer Interaction (HCI), Web of Science (WoS).

In the light of above table, it can be observed that reviewed literature mostly use English language resources to perform AKE. An Indian origin dataset is developed, but it also contains the news articles in English. In addition to these, there are several datasets reported in other languages that include Chinese, Japanese, Dutch, Korean and Turkish.

2.2.2. Summarization datasets

There exist multiple datasets for TS. Most of the datasets are constructed using newswire and scientific articles. Other sources that can contain ill-structured sentences include customer reviews from various resources such as eBay, Amazon and TripAdvisor. Following section compiles subset of the datasets that employed for TS.

2.2.2.1. CNN/DailyMail. CNN/DailyMail is one of the widely used dataset for TS. It was initially proposed for Question Answering (QA) Systems (Hermann et al., 2015) and later employed for summarization task (DMQA, 2019). First dataset is based on Cable News Network (CNN) that is an American based news channel. CNN dataset contains total of 90,000 newswire articles. Second dataset consists of 197,000 articles from Daily-Mail newspapers that is a United Kingdom newspaper. Each article in both datasets carries several bulleted points. Therefore, in order to use these datasets for summarization, respective bulleted points are treated as gold-standard summary against a newswire article. Studies employing this dataset include (Nallapati, Zhou, Gulcehre, & Xiang, 2016; Paulus, Xiong, & Socher, 2017; Rekabdar, Mousas, & Gupta, 2019; Song, Huang, & Ruan, 2019).

Articles against CNN were collected starting from April 2007 until the end of April 2015. Whereas for Daily Mail, data collection spans June 2010 till end of April 2015. In the rest of this study, this article is referred as CNN/DM.

2.2.2.2. GigaWord. GigaWord dataset (Graff, 2002) is also extensively used to evaluate TS approaches. It consists of vast collection of English newswire articles. Major four sources included in this dataset are listed as follows:

1 The Xinhua News Agency English Service
2 The New York Times Newswire Service
3 Associated Press Worldstream English Service
4 Agence France Press English Service

Amongst these various data sources, the first news source is included in dataset for evaluation of techniques on relatively noisy data; as this dataset used Machine Translation to translate original French news into English.

2.2.2.3. New York Times. Another summarization datasets named New York dataset, consists of millions of newswire articles from New York times (Sandhaus, 2008). The corpus includes over 1.8 million articles, where 650,000 article summaries are manually generated. In addition to summarization, this corpus contains tagged data for NER and RE. It also provides Java tools in order to parse the provided files.

2.2.2.4. Document Understanding Conference (DUC). It is one of the earliest workshops dedicated to the task of TS. It was organized from 2001–2007. After that, it is made part of Text Analytics Conference. The major dataset used in preparation of DUC are part of Topic Detection and Tracking (TDT) (Cieri et al., 2002) and Text Retrieval Conference (TREC) (Harman, 1996) collections, which

Table 2
Brief of datasets used for AKE.

Study	Genre	Count	Topic	Source	Language
Kim et al., 2010	Conference & Workshop	–284	DS, IR, AI	ACM	English
Inspec (Hulth, 2003)	Scientific Articles	2000	Computer Science and Information Technology	Inspec Database	English
Krapivin, ACM (Krapivin & Marchese, 2009)	Scientific Articles	2000	Computer Science	ACM	English
Aquino et al. (2014)	Scientific Articles	43	Computer Science	WICC	English
		166			
		72		CACIC	
MPQA (Rose et al., 2010)	News articles	535	Events and Opinions	CERATOPS	English
CSTR (Witten et al., 1999)	Technical Reports	1800	Computer Science	NZDL	
eBooks (Huang et al., 2006)	English books	101	Various Fields	–	English
Pay (2016)	Journal papers	1080	–	ACM	English
Turney (2000) (Used in El-Beltagy & Rafea, 2009)	Journal Articles	75	Various Fields	–	English
	Email Messages	311	–	–	
	Web Pages	90	–	Aliweb	
	Web Pages	141	–	NASA	
	Web Pages	35	–	FIPS	
Kumar and Srinathan (2008) *	Full-Length text document	1454	Various Fields	CogPrints	English
		104	Human Factors	ACM	
	Abstract Documents	34	HCI	CogPrints	
		194	Neuroscience		
		400	Philosophy		
Kumar and Srinathan (2008) (Used in El-Beltagy & Rafea, 2009)		80	CSTR	KEA 5.0	
UN-AF (Medelyan & Witten, 2006)[a]	Full-text Articles	200	Food and Agriculture	UN AFO	English
MAS (Yang et al., 2017)	Sentences/Words	900/28,500	–	MAS	English
Sterckx et al. (2018) *	Online News	1200 to 2000 Each	Online News	VRT	Dutch
	Online Sports		Online Sports		
	Lifestyle Magazines		Lifestyle, Fashion and Health	Sanoma	
	News articles		News articles	Belga	
El-Beltagy and Rafea (2009)	Journal 2	60	Various Fields	ACM, CogPrints	English
	Journal 3	21			
Song et al. (2017)	Tweets	200	Novelist	Twitter	Korean
	Tweets	200	Comedian		
Lahiri et al. (2017) [b]	SingleEmails	212	Informal & Formal	Enron collection	English
	Email Threads	107			
Zhang et al. (2017)	Scientific Articles	425	Computer Science	ACM WWW	English
	Scientific Articles	365		ACM KDD	
20News[c]	Newsgroup documents	18,845	Various topics	–	English
OMD[d]	Tweets	3269	Various government oriented topics	–	English
Jo (2003)	News articles	900	Nine various Fields	–	English
Wang et al. (2006)	Scientific Articles	250	Various Fields	Google Search	English
		50		Baidu Search	Chinese
Merrouni et al. (2016)	Journal Articles	82	Economics	Various Journals	English
		53	Legal (Law)		
		75	Medical		
Zhang et al. (2016)	Tweets/ Tweets_HT	1000	Randomly selected tweets	–	English
Marujo et al. (2015)	Tweets	1827	–	From October 27, 2010	English
Li, Du, and Xing (2017)	Abstracts	4272	Scientific Projects	NNSFC	Chinese
Park et al. (2014)	Blog Posts	14,000	–	Technorati	English
INND (Thomas et al., 2016)	Indian E-Newspaper	600	Various Topics	Times of India	English
		589		The Hindu	
		612		Hindustan Times	
		570		Indian Express	
Meng (Meng et al., 2017)	Scientific Articles	567,830	Various Topics	ACM, WoS, Wiley	English
KP20k (Meng et al., 2017)	Abstracts	20,000	Computer Science	ACM, WoS, Wiley	English
Disaster Tweets (Ray Chowdhury et al., 2019)	Tweets	122,266	Boston Bombing, Hurricane Harvey, and Hurricane Sandy.	Twitter	English

[*] Asterisks denote that respective datasets are available on request.

[a] www.fao.org/documents/

[b] http://lit.eecs.umich.edu/downloads.html#7

[c] http://qwone.com/~jason/20Newsgroups/

[d] https://bitbucket.org/speriosu/updown/src/1deb8fe45f60/data/shamma/orig/debate08_sentiment_tweets.tsv

consist of various topics. Amongst various genres, DUC majorly employs following datasets.

- TDT collections use news data from AP newswire and New York Times newswire,
- TREC collection whose subset is used, widely known as AQUAINT corpus (Graff, David, 2002) comprises:
  - ○ Xinhua News Agency (English version), 1996–2000
  - ○ New York Times newswire, 1998–2000
  - ○ AP newswire, 1998–2000
- TREC collection denoted as TRECNT
  - ○ Los Angeles Times 1989–1990
  - ○ Federal Register, 1994
  - ○ Financial Times of London, 1991–1994
  - ○ FBIS, 1996

Subset of studies performing summarization evaluation using DUC datasets include (Cao, Li, Li, Wei, & Li, 2016; García-Hernández & Ledeneva, 2009; Litvak & Last, 2008; Mihalcea, 2004; Ozsoy, Alpaslan, & Cicekli, 2011; Wang et al., 2018; Wong, Wu, & Li, 2008). Brief of subset of DUC tasks is shared below.

DUC 2001 makes use of NIST data carrying English news articles from various newspapers including Wall Street Journal, LA Times, Agent France Newswire and more. This data consists of thirty training clusters. Where, each of the cluster carries around three to twenty documents focused on a similar topic. On average, each cluster roughly carries ten documents with each having at least ten sentences. There exists summary against each document spanning 100 words. As this dataset also offers multi-document summaries, hence, each of the thirty clusters have four different summaries of length 50, 100, 200 and 400 words. Like the training dataset, the test data also carries thirty clusters.

DUC 2002 uses same document sources as that of DUC 2001. Total 533 news articles are used in this dataset for single-document summarization. In order to provide multi-document summaries, these 533 articles were divided into 59 clusters each having 5 to 15 documents. Furthermore, after pre-processing, each of the document is divided into three parts that include title, abstract, and main text. DUC 2002 also offers multi-document summarization. Hence four different multi document abstract summaries consisting of 10, 50, 100 and 200 words are also provided to evaluate multi-document summarization. Similarly, DUC 2003 consists of news from AQUAINT, TDT and TRECNT and also provides varied length summaries. Rest of the tasks makes use of AQUAINT corpus for document selection.

2.2.2.5. *Opinosis*. This dataset is based on customer reviews that are acquired from various websites related to different topics. Examples include reviews that rate quality of an electronic appliance, hotel ambience etc. Total of 51 various topics are part of the dataset. Websites that are used to acquire the customer reviews include Edmunds.com (cars), TripAdvisor (hotels), and Amazon.com (various electronics). Each topic in the dataset further carries 100 documents each. This dataset is employed in Ganesan, Zhai, and Han (2010) and Kågebäck, Mogren, Tahmasebi, and Dubhashi (2014).

2.2.2.6. *Summary*. As aforementioned datasets highlight the subset of datasets used in the literature for the task of TS. Therefore statistics against aforementioned and various other datasets that are employed by the studies covered in literature are presented in Table 3. Following nomenclature is used to describe the dataset. *Study* column describes the name of the dataset or the study in which the respective dataset is proposed. Genre column denotes the category of respective dataset. Numeric in parenthesis in this column represents length of human-annotated summary. C and D notations in *Count* column refers to number of clusters and documents respectively. E.g. “30C, 3-20D” that corresponds to first entry in *Count* column denotes that respective dataset consist of 30 clusters. Furthermore, each cluster has varied length ranging three to twenty documents. *Topic/Nature* either refers to the topics included in the dataset or the underlying nature of included text such as formal, semi-formal or informal. *Source* highlights the major sources used to develop the dataset whereas the primary language used in dataset is represented in *Language* column.

In the light of above table, it can be observed that reviewed literature mostly use English language resources to perform TS. In addition to these, there are several datasets reported in other languages that include Chinese, Japanese and Turkish. In addition very short datasets consisting of three and thirty documents are also used for low-resource languages such as Bengali and Hindi (Abujar, Hasan, Comilla, Shahin, & Hossain, 2017; Gupta & Garg, 2016).

2.3. Conclusion

Both evaluation metrics and underlying datasets used to evaluate proposed technique play a critical role in any machine learning domain. AKE and TS share some evaluation metrics that include precision, recall and f-score (P/R/F) which are mostly reported in AKE studies. On the other hand, for TS, ROUGE metric is widely used, which by its definition is a recall oriented metric. However, there exist many variations of ROUGE that enable reporting of TS results in form of P/R/F. In addition to these metrics, mean average precision, average match, error-rate and accuracy are some of the other metrics used to evaluate such systems.

As far as datasets are concerned, majority of existing datasets are based on news and scientific articles. These text genres comprise lengthy and well-formed sentences. Hence, sentence structure and various other linguistic features can be applied during processing of such data. In comparison to these, other datasets are focused on relatively short sentences that can contain formal, semi-formal or even informal text. Two of the most well-known data genres in this regard include tweets and emails. As such datasets can carry

Table 3
Brief of datasets for TS.

Study	Genre	Count	Topic/ Nature	Source	Language
DUC 2001	News Articles (50)(100)(200)(400)	30C, 3–20D	General	Various news sources	English
DUC 2002	News Articles (10)(50)(100)(200)	533, 59C, 5–15D	General	Various news sources	English
DUC 2003	News Articles(10) (100)	30C, ~10D each	General	AQUAINT	English
				TDT	
		30C, 22D	General	TRECNT	English
DUC 2004	News Articles (10)(100)(250)	50C, 10D	General	AQUAINT	English ~ translated from Arabic
		50C, 10D	General	TDT	
		25C, 10D		AFP Arabic NewsWire ~	
DUC 2005	News Articles (250)	50C, 25–30D	General	LA Times, FTL	English
DUC 2006	News Articles (250)	50C, 25D	General	AQUAINT	English
DUC 2007	News Articles (250)	50C, 25D	General	AQUAINT	English
	News Articles (100)	10C, 25D			
NYT (Sandhauts, 2008)	News Articles	1855,658	General	New York Times	English
CNN/DM (Hermann et al., 2015)	News Articles	312,085	General	CNN, DailyMail	English
SKE (Loza, Lahiri, Mihalcea, & Lai, 2014)	Email	349	Formal, Semi-formal, Informal	Enron	English
BC3 (Ulrich, Murray, & Carenini, 2008)	Email Thread	40	Semi-formal, Formal	Mailing Lists	
Opinosis (Ganesan et al., 2010)	Reviews	100	Hotels, Cars, Electronics	TripAdvisor, Amazon, Edmund	English
Yu and Ren (2009)	Web Pages	200	–	–	Chinese
	Articles	120	–	NTCIR	Japanese
Barzilay and Elhadad (1999)	Magazines	30	Economist, Scientific American	–	English
Reeve et al. (2006)	Clinical Trial Texts	24	Oncology	Drexel University College of Medicine	English
Lynn et al. (2018)	News Articles	600	General News, Technology	NYT, CNN, BBC, TechCrunch, Mashable	English
Sarkar (2012)	Bengali Documents	38	–	–	Bengali
Wu et al. (2015)	Single Documents	100	–		
Ozsoy et al. (2011)	News Articles	120	General	–	Turkish
	Scientific Articles	50	Medicine, Sociology, Psychology		
	Scientific Articles	50			
	Scientific Articles	153	Various Fields		
MatBN (Wang et al., 2005)	Broadcast News	4100	Weather Forecast, Advertisements, Headlines	Public Television Service Foundation[a]	Mandarin Chinese
Ma et al. (2018)	Reviews	34,686,770	Movie & TV	Amazon SNAP Review dataset[b]	English
			Sports		
			Toys & Game		
Collins et al. (2017)	Open-Access Articles	10,000 +	Various Fields	ScienceDirect	English
Khatri et al. (2018) [c]	Reviews	20,0000	Various Fields	eBay	English
Wang et al. (2017)	Conference	180	Various Fields	TIPSTER	English
	Journal	511		Various Journals	
	Legal Cases	1023		–	
	Patent	1021		USPTO	

[a] http://www.pts.org.tw.

[b] http://snap.stanford.edu/data/web-Amazon.html.

[c] https://github.com/EdCo95/scientific-paper-summarisation.

informal text and are relatively short in nature, therefore models exploiting sentence structure from predecessor might not work in successor. One solution to mitigate this challenge is to design robust algorithms that do not rely on underlying features of data at hand. Other option can be to train models on hybrid dataset, so that they can learn features of both and be able to distinguish them as well.

Furthermore, count of available datasets has increased over the years. This increase is primarily due to invention of crowd-sourcing platforms and improved annotation tools, which have improved the process of manual annotation in recent years. Also, the application of semi-supervised learning to annotate the data using a small portion of human-annotated dataset is another growing trend in dataset development process. Quality of such datasets can become a concern, so various solutions are proposed in literature to mitigate this issue. Nowadays, large datasets are being developed using crowdsourcing and machine learning techniques for AKE. As large sized datasets are required to effectively train deep neural networks. Hence, to combine various data sources to develop large datasets is also carried out in few studies, particularly for AKE.

3. Methodology

In order to collect the papers for AKE and TS, research articles having relevant sub-domain name in their titles were filtered using online computing research repositories such as IEEE, ACM, DBLP and Google Scholar. The selection methodology is loosely based on PRISMA guidelines that offer four-stage process for a systematic literature review. Identification phase starts with filtering articles using phrase searching was within titles to acquire the relevant literature. The basic set of keywords include: ‘Keyword Extraction’, ‘Keyphrase Extraction’, ‘Survey Generation’ and ‘Text Summarization’. Screening phase dealt with removal of duplicate articles along with those that were focused on other genres other than text, such as videos, after going through abstract and keywords. Next, eligibility phase filtered all articles that were available in full-text. Here, as generic text was being considered, total number of

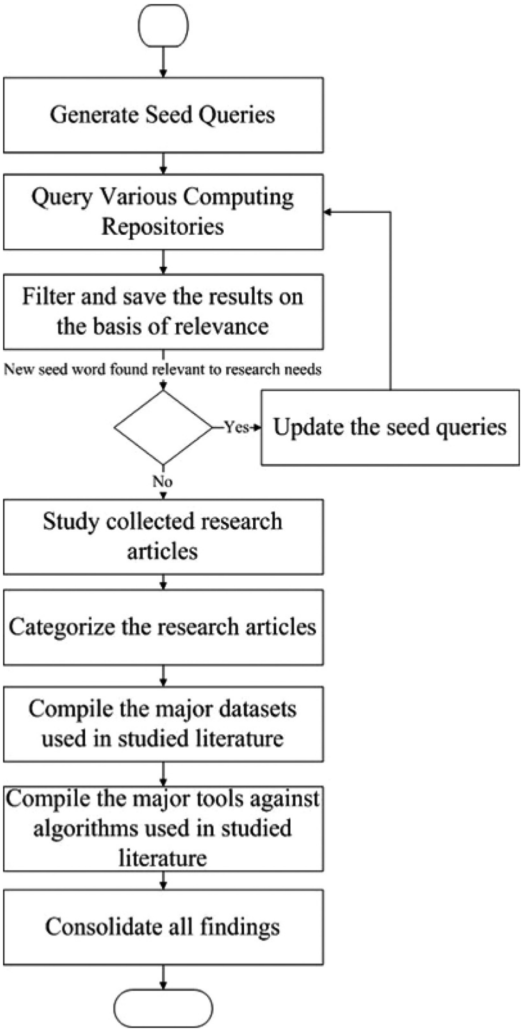
Generate Seed Queries
Query Various Computing Repositories
Filter and save the results on the basis of relevance
New seed word found relevant to research needs
Yes
Update the seed queries
No
Study collected research articles
Categorize the research articles
Compile the major datasets used in studied literature
Compile the major tools against algorithms used in studied literature
Consolidate all findings

Fig. 1. Complete flow of study.

Table 4
Papers distribution over the years included surveys for KE.

	Pre 2000	2001–2005	2006–2010	2011–2015	Post 2015	Total
Heuristic	1	3	2	1	3	10
Supervised	2	1	7	3	4	17
Unsupervised	0	1	4	6	9	20
Neural Network	0	1	1	3	3	8
Survey	0	0	0	4	1	5
Total	3	6	14	17	20	60

relevant studies after screening were quite high. Therefore, while reading full-text, representative studies against each approach are selected, for the sake of brevity, to present overall progress over the years. In addition, the primary focus of this search study is to highlight contributions made with respect to deep learning frameworks, as existing survey do not cover them extensively. As deep learning approaches can be regarded as multi-layered neural networks, having more than one layers of hidden neurons. Therefore neural network based approaches are covered extensively.

Fig. 1 shows complete flow chart, containing all major tasks performed to conduct this study. Initial queries resulted into approximately three hundred studies. Resultant set was later manually pruned to filter studies based on underlying approach. Final count of selected studies for AKE and TS was 60 and 51 respectively. The yearly distributions of studies for AKE and RE are presented in Table 4 and Table 5 respectively. For brevity, these are grouped into span of five years, starting from year 2000 to 2019, where earlier studies from 2000 are being catered as Pre-2000 and Post 2015 refers to papers that are published after 2015. Table statistics show that even now, supervised and unsupervised solutions are being used for AKE. On the other hand, for TS, application of neural network based deep learning approaches has increased in recent past. Rest of the paper describes AKE and TS in detail. Each section covers existing survey studies, current state-of-the-art using deep learning and other technologies followed by conclusion.

4. Keyword extraction

AKE is being widely addressed in literature, where keywords can be single word or multi-word (in the form of phrases). Following sections describe the existing survey studies in context of AKE followed by the classification of existing literature. As majority of the existing surveys classify on the basis of employed approaches. Therefore, current study also follows the similar classification.

4.1. Existing surveys

In this section, various survey studies that are focused on the task of AKE are briefly explained. These studies tend to cover variety of aspects including techniques, datasets, widely used features of sentences and words and various evaluation metrics.

A review study presented in Menaka and Radha (2013) has focused or classified AKE techniques into neural based approaches, lexical chain based approaches, graph based approaches and word co-occurrences based approaches for AKE. In addition to that, it compiles set of features that are being used against each approach as well. Study concludes that neural based approaches perform better than the rest.

A state-of-the-art exploration of AKE is also presented in Hasan and Ng (2014) which firstly provides brief overview about data features that result in the complications for AKE that include length, structural consistency, topic change and topic correlation. Secondly, it provides classification for major AKE approaches into heuristic, supervised and unsupervised approaches. Thirdly, it lists major errors that are currently being generated by widely used AKE systems where errors are broadly classified into precision and recall related errors. Precision related errors include over-generation errors and redundancy errors. Over-generation refers to errors generated when multiple keywords are selected as they tend to contain highly frequent potential word, whereas redundancy errors occur when a keyword is selected along with its semantically equivalent phrase or alias. Infrequency errors and evaluation errors are amongst the recall related errors. Infrequency errors occur due to infrequent occurrence of keyword in its respective document, whereas evaluation errors are made when evaluation system is unable to semantically compare the similarity amongst the identified keyword with the ground truth. Finally, authors give their opinion about the current trends and challenges regarding AKE. it is recommended to use external resources such as DBPedia and Freebase to improve the results and avoiding the listed errors.

Table 5
Papers distribution over the years included in survey for TS.

	Pre 2000	2001–2005	2006–2010	2011–2015	Post 2015	Total
Heuristic	1	0	0	1	3	5
Supervised	0	0	3	0	0	3
Unsupervised	2	3	3	7	6	21
Neural Network	0	1	0	2	15	18
Survey	0	0	1	1	2	4
Total	3	4	7	11	26	51

Evaluation errors can only be avoided by means of improving evaluation algorithms, as this error is more related to the evaluation system. Thus, an evaluation criterion that could compare various keywords keeping their semantics in view is of utmost importance. Amongst all the surveys reviewed, this survey stands out the most in terms of its contribution.

The survey study in Siddiqi and Sharan (2015) has its major focus on classification of approaches as well as features to perform AKE. Firstly, approaches are classified in four major classes that include statistical methods, supervised, semi-supervised and unsupervised approaches. Later, classification of features is carried out based on their phraseness and informativeness. The phraseness features represent how words in a phrase are connected to each other. In other words, it gives the probability of considering set of words as a phrase. Such features include mutual information as well as mean and variance. Informativeness features, on other hand, provide insight about the goodness of keywords/keyphrases. These features are further classified in three types that include features based on term weight, location and miscellaneous entities. In addition to that, this study divides the task of automatic keyword generation into two tasks namely keyword assignment and AKE. Here, keyword assignment refers to selection of keywords from predefined taxonomy or vocabulary with its major challenge being the construction of such resources. Similar division of keyword generation into keyword assignment and AKE is termed as 'automatic keyword indexing' in Beliga (2014) and (Meštrović) as well.

A survey study majorly focuses on graph based approaches for AKE in presented in Beliga et al. (2015) . This study highlights the major advantages of graph based approaches that include language-independence, linguistic knowledge independence and domain independence. Study covers various graph types with respect to representations of vertices and edges. As a vertex can either represent concept or weight, further two classes of vertices representations are made. Similarly, edges based representations are divided into three groups as edges can be used to express direction, label or weight of the association. In addition to provide taxonomy for graph based algorithms, study also highlights widely used centrality measures to identify potential vertices within a graph using ranking of vertices. Various centrality measures are used to perform AKE which include TextRank, HITS, closeness, in-degree, and out-degree. Study has performed comparative analysis of various centrality measures that include in-degree, out-degree, closeness, betweenness and selectivity. Results evaluated on Croatian news dataset show that selectivity tends to provide better keywords whereas rest of the approaches selects stop-words as top-ranking keywords. Thus, study emphasized on employing and improving graph based approaches for NLP-oriented tasks.

Another study presented in Merrouni, Frikh, and Ouhbi (2016) divides the task of AKE in four major steps that include data preprocessing, candidate keywords selection, keyword ranking and evaluation. Data pre-processing include tokenization, lemmatization, POS-tagging and chunking of data. Candidate keywords selection is focused on selection of keywords on the basis of various text related constructs or heuristics that include n-gram based approach, chunking information and POS information. Keyword ranking can be carried out in either supervised or unsupervised fashion, where candidate keywords from previous phase are to be ranked. The study briefly explains supervised and unsupervised approaches in light of widely used AKE systems. Finally; in evaluation phase evaluation of automatically extracted key phrases in either manual or automatic fashion is performed. Manual evaluation involves humans to verify that either extracted keywords are relevant to document or not, thus it is quite expensive to perform. On the other hand, automatic evaluation tends to compare extracted keywords with ground truth by means of precision, recall and f-score. In addition to that, this study provides directions for future work that include improvement in pre-processing states and candidate extraction phase. The study also highlights the importance of feature selection as it tends to heavily affect the performance of supervised approaches. Furthermore, the study recommends use of external knowledge bases as these external resources help in better understanding of keywords. Lastly, authors have focused on semantic based evaluation of keywords extraction approaches as previously highlighted in Hasan and Ng (2014).

In the light of existing surveys, literature is classified using widely used approaches that include heuristics, supervised and unsupervised approaches. That's the primary motivation to categorize reviewed literature in current study on the basis of underlying approaches. Moreover, none of the reported surveys cover applications of recent deep learning approaches in the area of AKE. Hence, this is the primary distinction of this study. Table 6 highlights the major emphasis and features of existing surveys. In this table PR denotes whether process and repositories to acquire literature is described. HB, SU and US denotes whether survey included heuristics based approaches, supervised approaches, and unsupervised approaches classification. Lastly, ST denotes whether summary table of reviewed literature is presented in a survey. Value CI denotes comprehensive information, whereas PI denotes partial information.

4.2. Heuristic based approaches

Research studies in this category usually rely on two types of heuristics, where one deals with selection of keywords, while other is concerned with pruning of words/phrases that are not fit to be primary keywords/keyphrases. Some of the key-heuristics include

Table 6
Summary of existing AKE surveys.

Study	Major Focus/ Distinctions	Coverage	PR	HB	SU	US	ST
Menaka and Radha (2013)	Graph-based, Lexical chain, MLP, Word co-occurrence	1998–2010	–	–	–	–	PI
Hasan and Ng (2014)	Features of datasets, Various types of errors, Challenges in AKE	1998–2013	–	CI	CI	CI	PI
Siddiqi and Sharan (2015)	General	1975–2012	–	CI	CI	CI	–
Beliga et al. (2015)	Graph-based methods, Croatian language	1998–2014	–	PI	PI	PI	CI
Merrouni et al. (2016)	General	1999–2015	–	–	–	–	CI
This study	Neural network approaches, underlying datasets	1998–2019	CI	CI	CI	CI	CI

exploitation of POS-tag such as noun, adjective, selecting words from external resources as candidate words or extracting terms following a pre-defined pattern. Heuristic based approach for AKE proposed in Wu, Li, Bot, and Chen (2005) makes use of keyphrase identification by means of human annotated samples. First a glossary is built using human annotations which is followed by identification of noun phrases amongst them. Another threshold-based approach that employs n-grams for multi-lingual keyword extraction is presented in Tseng (1998).

A study to perform AKE using pattern filtration is presented in Kumar and Srinathan (2008). First, data is compressed using Lempel–Ziv compression algorithm LZ78 (Kagan, 2019) to construct n-grams. These n-grams are later refined using simple rules that take into account POS information. In the end heuristics and lexical information is used to extract keyphrases.

Study presented in Pay (2016) makes use of four-step process to perform AKE. The algorithm is named "TAKE", and the steps include candidate extraction using noun-phrase followed by filtering using location heuristics. After that, various word frequency and its degree in candidate feature set is used to calculate keyword score. This step is followed by a dynamic selection of threshold using either mean or medians of all candidate keywords. Results are compared with Rapid Automatic Keyword Extraction (RAKE), Rose et al. (2010) and TextRank (Mihalcea & Tarau, 2004). Later, a follow-up study (Pay & Lucci, 2017) that employs an ensemble method using TAKE, RAKE and TextRank is also presented which outperforms each individual component.

POS information, lemmatization and named entity recognition are being used to extract keywords from tweets in Weerasooriya, Perera, and Liyanage (2017). Proposed approach starts with development of domain specific corpus. This corpus contains manually collected 206 words after analyzing tweets. Using this corpus and various other features, AKE is performed. Results are evaluated on fourteen tweets using Turing test i.e. human annotated phrases against these fourteen tweets are compared against automatically extracted keywords. The results show that proposed approach resulted in P/R/F of (72.0, 91.0,78.0). Heuristics are usually applied along with supervised and unsupervised approaches for AKE to improve the results. The studies carried out in Hulth (2003), Mihalcea and Tarau (2004), Newman, Koilada, Lau, and Baldwin (2012) and Wan and Xiao (2008) makes use of POS-tag related heuristics to perform AKE.

4.3. Supervised approaches

A widely used algorithm for AKE is KEA (Witten et al., 1999) algorithm that takes into account the features of the words such as word's TF-IDF along with its occurrence information. It later employs Naïve-Bayes classifier for keyphrase extraction. The proposed approach was evaluated on CSTR by NZDL. Total of 1800 reports were used where author has already listed the keyphrases against respective report. Out of these 1800 reports, 500 were used as testing dataset. The study experimented with various parameters including total number of keyphrases to be extracted as well as impact of training set size on overall results. Average number of matches is used as evaluation measure between extracted keywords and author-assigned keywords. An extension to KEA algorithm named as KEA++ (Medelyan & Witten, 2006) is also proposed that incorporates domain-specific thesaurus. It uses basic KEA-features along with node degree that uses domain-specific thesaurus information and length feature. Other feature based approaches to use SVM, Naïve-Bayes and random forest are reported in Nguyen and Luong (2010), Treeratpituk, Teregowda, Huang, and Giles (2010), and Wu, Marchese, Jiang, Ivanyukovich, and Liang (2007) respectively.

Another study for AKE that takes into account various linguistic features is presented in Hulth (2003). It majorly provides a comparison of incorporating linguistic features such as syntactic tags using Noun-Phrase chunking, and POS tags in the overall performance. To evaluate the results, dataset for Inspec database is used. In order to train the model, bagging ensemble is used with all features. Experiments show that incorporation of linguistic feature results in improved performance in comparison to relying only on frequency oriented features.

HUMB (Lopez & Romary, 2010b) is another key-extraction system that is tailored for scientific articles. It uses supervised algorithms after making use of structure based features using GROBID (Lopez, 2009). GROBID is a java-based solution to extract and parse scientific articles in xml format. After primary processing using GROBID, HUMB further extracts concept based features using Wikipedia and GRISP (Lopez & Romary, 2010a), where GRISP is terminology database for scientific articles. Decision Trees, Multi-Layer Perceptron and SVM are used for classification along with bagging and boosting techniques. As per the result, bagged decision trees performed better than the rest with SVM resulting into the least accuracy.

A research carried out in Huang, Tian, Zhou, Ling, and Huang (2006) view each document as a semantic network, where each network node contain syntactic and statistical information. After performing summarization on all nodes, nodes are further filtered and scored. Based on scoring, keyphrases are extracted. In order to further improve the accuracy, SVM classifier is also trained to dynamically predict the number of keyphrases as per the input. Experiments show the accuracy up to 80%.

Supervised algorithms for keyphrase extraction are also used in Turney (2000) . This study firstly used Decision Tree induction algorithm with total of nine (9) features that includes word stem, word count, occurrence, POS information and finally its class (is it a keyphrase or not). Testing was based on similarity of human-assigned keyphrases with extracted ones. Another algorithm for keyphrase extraction was proposed in this study that is based on genetic algorithms. This genetic algorithm based keyphrase extractor is termed GenEx. Experiments conducted via GenEx shows that it performs better than Decision Tree algorithm as GenEx is tailored primarily for keyphrase extraction. SZTERGAK (Berend & Farkas, 2010) is another supervised system that takes into account many features at different levels including phrase, document, corpus and knowledge base. The proposed system performed well in shared task and stood third overall.

A research study (Haddoud & Abdeddaim, 2014) proposed measure for AKE known as Document Phrase Maximality Index (DPM-index). DPM-index is introduced in order to measure the overlapping between candidate keyphrases. The study employs supervised approach to AKE by making use of eighteen statistical features that include frequency, length, likelihood estimates and divergence related features along with DPM-Index. Resultant system outperformed existing best performer of Sem-Eval AKE task. P/R/F on extracting top five, ten and fifteen phrases are (44.8, 15.3, 22.8), (36.2, 24.7, 29.4) and (28.3, 28.9, 28.6) respectively.

Another study that majorly uses keyphrase extraction in order to organize a document summary in survey article is presented in Yang et al. (2017) . This study uses Conditional Random Fields (CRF) in order to identify potential keyphrases. This study primarily takes citation sentences of a document as an input. CRF is applied on set of citation sentences in order to extract potential keyphrases. Total of twenty-two (22) features are used in order to train CRF. These features include word window of two, POS information, term frequency, casing information, mutual information, correlation features, punctuation information, relation information and some combination of above mentioned features. Amongst these features, relation and POS information is acquired using Stanford Core-NLP (Manning et al., 2014) that is trained over Penn-Tree Bank dataset. Feature related to mutual Information is evaluated using English Wikipedia corpus (Wikipedia, 2019). The citation dataset was taken from Microsoft Academic Search (Microsoft, 2019) for evaluation. CRF shows great improvement with F1-measure of 70.5 when compared with existing algorithms including KEA and SVM with F1-measures of 49.9 and 64.6 respectively.

An approach for AKE and TS is presented in Thomas et al. (2016) on Hindi e-newspapers. It employs HMM for POS tagging. Using these tags and count of each individual word, probability of words w.r.t POS tags are calculated. This information is later used to perform AKE. Using these keywords, a TS algorithm is also proposed that allows users to input desired number of sentences in summary. Using this number, authors have employed various ranking schemes including first-come, first serve and TextRank.

Another algorithm is proposed in order to extract document-specific keyphrases by means of sequence mining (Xie, Wu, & Zhu, 2017). It majorly consists of three phases namely pre-processing, pattern-matching followed by classification. This study takes collection of words as a word sequence thus; sequential pattern mining is employed along with wild-cards in order to extract potential document-specific keyphrase. Patterns are processed in depth-first fashion so that prefix pattern can be reused. After acquiring potential candidates for keyphrase, the results are later classified using Naïve-Bayes classifier, which helps in classifying keyphrases with respect to the document content. In order to train the Naïve-Bayes classifier, total of six features are employed where two of them are regarded as base-line features that include TF-IDF and position of first-occurrence of respective word. Remaining four features are related to extracted patterns. After classification, top-k phrases are returned as keyphrases. Experiments are performed on Reuters-21578 collection that carries manually labeled newswire articles with total of 135 categories. 'Acq' category is used in this study for evaluation purposes. Another dataset used is SemEval-2010. Results show that proposed algorithm outperforms the previous algorithms KEA (Witten et al., 1999), and KP-Miner (El-Beltagy & Rafea, 2009) in terms of F1-measure with max-score of 19.77. It is computationally efficient than the previous algorithms as well.

A comprehensive experimental study is carried out in Sterckx et al. (2018) that employs various algorithms to a newly constructed dataset. Four major Dutch language data sources are used for annotation; that include news, movies, life style magazines and printing press. Annotation guidelines and an annotation tool is also developed to assist during manual annotation. On annotated dataset, results are reported using variety of supervised and unsupervised approaches. Results show that gradient boosted decision trees outperform various other approaches when background knowledge is incorporated using Dutch Wikipedia. Other supervised approaches for AKE include Caragea, Bulgarov, Godea, and Gollapalli (2014) and Haddoud, Mokhtari, Lecroq, and Abdeddaïm (2015).

4.4. Unsupervised approaches

Researchers have employed various unsupervised approaches to perform AKE in previous years. TextRank algorithm is amongst the pioneer studies to perform AKE in unsupervised fashion (Mihalcea & Tarau, 2004) . This study makes use of graph based approach along with POS tag information to perform AKE. In order to evaluate, Inspec dataset was used which resulted into f-score of 36.2. Another AKE algorithm named "RAKE" employing TextRank scores is proposed in Rose et al. (2010). Apart from TextRank, two major modules are involved to improve AKE. First one is stop list generation that takes manual input to define the list of words that must be excluded from candidate keywords. Consequently, it can result in loss of meaningful phrases, carrying the stop words. Hence, the second major module is focused on word adjoining. This is achieved by looking for at least two occurrences of adjoined keywords within a document. The approach is evaluated on Inspec and has resulted in improved performance over the existing techniques. In addition, top ranking keywords against MPQA corpus are also extracted and results show that proposed approach performs better.

Another graph based approach presented in Wan and Xiao (2008) makes use of neighborhood knowledge to incorporate the global information as well. Firstly, neighborhood documents are added into set using document similarity techniques. Later graph based ranking approach is applied on resultant set in order to perform AKE. This approach exploits local information contained in paper as well as global information that is collected using neighborhood papers as well. Tagged dataset for AKE was prepared using DUC 2001 dataset for summarization. Experimental results show that proposed approach outperformed various state-of-the-art approaches for AKE.

KP-miner is another unsupervised keyphrase extractor (El-Beltagy & Rafea, 2009). It basically focuses on understanding and exploitation of keywords identification process. Thus, this approach employs heuristics as well as feature, where most of its features are related to the nature of general keywords and documents, rather than any restricted domain. The overall system consists of four phases. First phase is candidate keyword selection which deals with selection of candidate keywords on the basis of heuristics related to punctuation marks, position and stop words. Second phase deals with weight calculation of candidate keywords by means of TF-IDF, boosting factors as well as position features. Final phases deal with refinement of keywords that return required number of keywords based on calculated weights. Experiments are conducted on English and Arabic languages, where framework is extensible to any other language. Results show that KP-miner outperformed KEA in all scenarios.

Another approach that employs PageRank to perform AKE is presented in Liu, Huang, Zheng, and Sun (2010). As graph oriented approaches tend to highly rank the words that are strongly connected. Hence, they might end up missing some relatively sound keywords. In order to deal with this, authors apply intuitive idea to start multiple PageRanks tailored to a specific topic. Hence, this approach is called Topical PageRank (TPR). The process first starts with topic discovery from large collection of words using Latent Dirichlet Allocation (LDA). After that, a word graph is constructed and TPR is employed using various topics discovered in previous stage. Results are then ranked and top occurring keywords in terms of frequency are selected as final set of keywords. Experiments are conducted on 308 news articles from DUC 2001 and Inspec dataset. Experiments are compared against various solution including PageRank, TFIDF and LDA as well. Results show that proposed approach outperformed all against both datasets.

Research study presented in You, Fontaine, and Barthès (2013) uses core word set. Core words are the most frequent words occurring in a document that are non-stop words and have different stems. After extracting core word set against a given document, this study uses position features and statistical features such as frequency, length in words, and core word related feature to find out candidate keyphrases. Lastly granularity-related features are employed to eliminate any overlapping phrases. Overlapping phrases refers to those when a phrase as well as its sub-phrase, both are part of candidate set. For example, one phrases in candidate list is ‘network optimization’ and other one is ‘Semantic network optimization’ then such phrases are called overlapping phrases. Which of the overlapping phrases should be part of the final set is governed by the granularity features. Once feature values are learnt, respective parameters are robust enough to be applied as it is on unseen data. Results are evaluated on Sem-Eval task dataset as well as self-generated dataset titled ‘VarTypes’ which contains more diversified scientific documents than the available data. When compared with other techniques of Sem-Eval task, this study stood second with HUMB being the best keyphrase extractor. The research proposed in Park, Kim, and Lee (2014) is focused on keyword extraction from blogs based on content richness. This study primarily proposes a new evaluation measure to extract keywords that is named ‘richness’. Richness measures how much coverage is provided in a blog against a keyword. This research also makes use of external data to improve the richness aspect. Outside data is gathered by means of querying candidate keywords over Google search engine and collecting sub-topics against each keyword. In order to conduct experiments, fourteen blogs from Technorati were selected and keyword set was manually prepared for evaluation. The results were then calculated against proposed approach and various other unsupervised models including TextRank, TF-IDF and HITS. Comparative analysis shows that the proposed approach out-performed other unsupervised approaches for AKE.

A research study focuses on extraction of contextual keywords by means of crowd-sourcing (Hong, Shin, & Yi, 2014) . The proposed approach in this study consists of four steps that include frequency based term extraction, sentence construction, re-weighting of terms based on constructed sentences and finally sentence voting. First step deals with filtering on the basis of POS-tags as well as frequency. Second step involves human users to select sentences that best describe the document, using information acquired in first step. Later, in third step, terms are reweighted based on the human-selected sentences. Last phase of sentence voting also involved users to rank the sentences. The study emphasis on employing sophisticated machine learning techniques along with human input by means of crowdsourcing. Such kind of hybrid approach can lead towards bridging the gap between human contextual understanding and AKE approaches. Results show that proposed approach resulted in precision and recall of 62% in comparison to simple frequency-based approach that resulted in precision and recall scores of 46% each.

A study reported in Beliga (2014) classifies existing literature in supervised, semi-unsupervised and unsupervised methods. In addition to that, it focuses specifically on Croatian language as well as selectivity based keyword extraction (SBKE) technique. SBKE is a graph based unsupervised approach which employs a novel measure for node selectivity value in order to perform AKE. The proposed selectively measure outperform existing graph-centrality measures when applied on Croatin news dataset resulting in F-score of 25.9. This approach is further applied in follow-up study as well (Beliga, Meštrović, & Martinčić-Ipšić, 2016) to improve the ranking phase of AKE. Major advantage of selectivity feature is that it does not require any linguistic knowledge.

A recent study carried out in Song, Go, Park, Park, and Kim (2017) focuses on just-in-time AKE from meeting transcripts. This problem is different from conventional AKE as new keywords are to be extracted just in time. This study takes into account three factors that include temporal history, topic relevance and participants. A graph based AKE is proposed to perform this task, where each graph node represents candidate keyword and edge represents their temporal importance. For history management, two types of separate history graphs are being maintained where one is dedicated towards content based information while other concerns about participant related history information. In addition to that, a forgetting curve (Wozniak, 1999) is employed to update the weights in history graphs in order to give more weight to recent activities. TextRank is used to extract keywords from both graphs and later result is aggregated by means of weighted sum in order to extract overall keywords. The experiments are conducted on ICSI corpus (Janin et al., 2003) as well as self-generated twitter data with average f-scores of 33.6 and 49.8 respectively outperforming other state-of-the-art systems including TF-IDF based solutions and TextRank system.

Research carried out in Lahiri, Mihalcea, and Lai (2017) focuses on keyword extraction from emails text. Major contribution of this study was dataset preparation for email text. This study made use of both supervised and unsupervised methods in order to identify keywords from emails. Amongst the supervised ones, KNN and Naïve-Bayes were used with different features that were related to individual words and phrases, whereas TF-IDF and neighborhood algorithms are used for unsupervised keyword extraction. Results are compared with other AKE systems including KEA, SZTERGAK, SingleRank and ExpandRank. Amongst the various algorithms for email keywords extraction, supervised KNN tend to perform better than all other systems with P/R/F of (31.94, 50.03, 38.99) which gave best results. Unsupervised algorithm employing mean neighborhood information performed better than existing unsupervised approaches for AKE with P/R/F of (27.33, 35.33, 30.82).

MIKE is an AKE system that integrates the features of both supervised and unsupervised into one solution (Zhang et al., 2017). It is focused on integrating widely used five features into a model that include word graph structure, topic information, features related to candidate words and links in a graph and prior knowledge. This ensemble of features has resulted into a unified model that captures the advantages of both approaches. Experiments are conducted on ACM WWW and KDD conferences, which resulted into P/R/F of

(14.8,15.1,14.9) and (16.1,16.0,16.0) respectively.

A rather recent approach (Alrehamy & Walker, 2018) that employs unsupervised clustering to perform AKE using WordNet as a knowledge base. It firstly identifies candidate keywords using WordNet along with employment of various NLP constructs such as POS tagging and chunking. Later word sense disambiguation (WSD) is performed using SenseRelate-Target Word method. Furthermore, an adjacency matrix is constructed that holds semantic similarity among candidate words using WuPalmer measure. This matrix is later clustered using Affinity Propagation that also produces exemplars against each cluster, which can be regarded as the semantics of respective cluster. Using these exemplars, seeds are selected and later pruned by employing set of heuristics along with a rule-base. This whole AKE process is named SemCluster. In order to evaluate the performance, Inspec and DUC 2001 datasets are used and comparative analysis is performed against TextRank, ExpandRank and KeyCluster. Results show that proposed SemCluster outperforms the rest with P/R/F of (40.1,74.2,52.0) and (36.4,69.2,47.7) on Inspec and DUC 2001 respectively.

Another approach exploiting topic modeling to perform AKE on social media text is presented in Yang et al. (2018). Process consists of three phases, with first one being phrase-mining using (El-Kishky, Song, Wang, Voss, & Han, 2014). Once phrases are extracted, next step is to perform topic modeling. This is carried out using an extension to LDA framework. The proposed extension named Task-oriented LDA (ToLDA) seeks input related to specific topic/event a user is interested in to better address the needs. In ToLDA, the latent variables are being learnt using Gibbs Sampling. After primary keyphrase extraction, last step deals with synonym identification using PMI-IR (Turney, 2001). Other unsupervised techniques for AKE include DegExt, SwiftRank and various other graph-based approaches (Abilhoa & de Castro, 2014; Litvak, Last, & Kandel, 2013; Lynn, Lee, Choi, & Kim, 2017; Tixier, Malliaros, & Vazirgiannis, 2016; Ying, Qingping, Qinzheng, Ping, & Panpan, 2017).

4.5. Neural network approaches

This section covers the research done for AKE by means neural network approaches. Widely used approaches in this regard include Multi-Layer Perceptron (MLP) and Recurrent Neural Network (RNN).

4.5.1. Multi-layer perceptron

A pioneer study to employ neural networks for AKE is proposed in Jo (2003) which majorly compares the result of proposed system with simple TF-IDF based approach. In this particular study TF-IDF is being termed as equation based approach. It uses term frequency, inverse document frequency, existence of phrase in title, existence of word in first and last sentences of document as input to neural network, whereas output layer has two nodes that outputs features related to keyword and non-keyword respectively. Comparison with simple equation based TF-IDF shows that neural networks with back propagation tend to perform better.

The study proposed in Wang et al. (2006) makes use of multi-layer perceptron in order to perform AKE. It uses term frequency, inverse document frequency, existence of phrase in title and phrase structural feature as input to neural network. Threshold of 0.5 is used at output layer to crisp classification of word into keyword or non-keyword. The set of 300 documents with two fifty containing English Text and remaining fifty containing Chinese text are processed manually and sample of 237 documents amongst them are annotated by author. By comparing system generated keywords with author-generated ones, resulting precision and recall are 31.6 and 37.5. Human based evaluation is also performed which results into 64.7 good phrases, 15.8 of fair keywords and rest proportion of poor keywords. Run-time experiments are also conducted which shows that proposed approach is computationally efficient.

The study in Sarkar et al. (2012) compares the employment of various ML-constructs including Naïve-Bayes, Decision Trees and Artificial Neural Networks for the task of AKE. This study uses dataset of 210 Journal papers in order to perform comparative study. Results show that MLP approach for Artificial Neural networks tend to outperform the other two classifiers.

The research presented in Aquino et al. (2014) deals with AKE problem from Spanish scientific articles. It treats AKE as a recognition problem rather than discrimination problem. Auto-encoder tends to learn the features of input as its expected output is basically the input passed to it. As NNs tend to map the input to output via feature learning, auto-encoders tend to map the input to input itself in order to learn the features of input. Input of model consist of eight features that are quite similar to the one used in Jo (2003). These features include normalized sentence length and normalized frequency as well. In order to train the network, resilient back propagation algorithm is used. Results are presented on three corpora consisting of scientific articles where two datasets consist of articles from Workshop of Researchers in Computer Science (WICC) and one dataset carries articles from Argentine Congress of Computer Science (CACIC). Two datasets from WICC carry 43 and 166 documents respectively whereas third dataset from CACIC consists of 72 articles. Results show that proposed system tends to outperform existing systems that include KEA and KEA + +. P/R/F against three datasets using proposed methodology are (36.2,34.3,34.4, 26.6,24.9,25.6) and (13.8,17.7,15.3) respectively.

Study proposed in Jo and Lee (2015) makes use of deep belief networks (DBN), that is a variant of MLP, where instead of communications between perceptron, layers are linked with each other. DBNs are used to extract latent keyphrases. These refer to phrases that do not occur in the document but are valuable for document insights. One-hot word vector representation for whole document is given as input to DBN. Output layer is a logistic regression layer which outputs candidate phrases. A weighted cost function is used in order to improve the extracted phrases. Results are evaluated on INSPEC dataset and compared with LDA based Latent key phrase extraction approach proposed in Cho and Lee (2015) which shows improvement in f-score as long as damping factor resides in the range [0.5–0.9]. The major idea of proposed approach was to learn the intrinsic features of documents and using them to find latent key phrases.

4.5.2. Recurrent neural network

Recurrent neural networks based approach is proposed in Zhang, Wang, Gong, and Huang (2016) on twitter dataset. This study is pioneer in applying deep learning solutions for AKE. As tweets are short in length, traditional frequency based features are not very suitable for this type of data. The study proposed a model consisting of two layer RNN that jointly models generation and ranking of keyphrases. One layer is used to discriminate keywords and second one is used to filter keyphrases. These two sub-objectives performed in different layers are later combined into final objective function. Stochastic Gradient Descent (SGD) is used to train the model. Word embeddings are used as input to RNN and also to initialize word weight matrix. Dataset was constructed by crawling and algorithmically assigning keywords on the basis of hash-tags. On the generated dataset, study compares the results of proposed approach with CRF, AKET (AKE for twitter) (Marujo et al., 2015) and other variants of RNN including LSTMs, Simple RNN and R-CRF (RNN + CRF) (Yao et al., 2014) models. For keyphrase extraction task, proposed algorithm results in P/R/F of (80.74, 81.19, 80.97) outperforming all other models.

In the aforementioned study, experiments are also performed with respect to hyper-parameters of RNN including number of neurons, window size, value of alpha which is a weighting factor between keyphrases and keywords objectivity in final trained model. In addition to that, various word embeddings based experiments are also reported. The proposal of joint model using RNN for automatic keyphrase extraction and ranking is the major contribution of this study. A rather recent study that perform AKE to improve the task of document classification is presented in Wu, Du and Guo (2018). It incorporates the concept of visual semantic annotation during the process of AKE by means of calculating log-likelihood as per the word count and normalization. Using this information, bi-directional LSTM is employed to classify the documents. Another approach employing bi-directional LSTMS to capture long-distance and CRF to model sequence labeling dependencies is presented in Alzaidy, Caragea, and Giles (2019). A similar approach that employ Bi-LSTMs with variety of word-embedding on twitter data is carried out in Ray Chowdhury, Caragea, and Caragea (2019) to perform AKE on disaster related tweets.

4.6. Conclusion

AKE is one of the major building blocks in NLP. It serves as baseline for many advanced NLP tasks including Question Answering Systems, Text Summarization and Text Similarity Analysis. Majority of the work currently focuses on standard language text that mostly belongs to scientific articles or news articles. In the light of current-state-of-the-art, among the existing approaches including machine-learning based supervised and unsupervised; graph based approaches are resulting into relatively better performance, in comparison to other approaches. Their primary advantage is that they do not employ any domain-specific, language-specific or linguistic knowledge.

Deep neural networks (DNN), on the other hand, are less applied for AKE in literature. RNN and its variant LSTM are among the widely used DL framework in AKE. However, there exist various other concepts such as CNN and joint modeling that are yet to be examined on AKE. Whereas, these techniques seem to perform well in other state-of-the-art text mining tasks such as Named Entity Recognition, Relation Extraction and Textual Summarization. Therefore, application of these approaches to determine their effectiveness in performing AKE is a research direction.

So far, the studies that have employed deep learning frameworks are focused on scientific articles and tweets. The primary reason for that is development and availability of datasets that are big enough to train the DNN. If these results are compared to existing graph-based approaches on similar nature of datasets, neural network approaches seem to perform at par and sometimes even better. However, the underlying datasets used in respective studies are not similar. In addition, as DNN requires huge data to train and it is a common knowledge that, increase in training data improves the performance of machine learning models. Therefore, if similar (huge) dataset is used to train a feature-rich/ sophisticated machine learning model, it might also result in performance improvements. Hence, a comparative study is required to analyze the effects of one approach over the other, on the available datasets. Such studies can further guide the direction and prospects in AKE.

To conclude the advancements regarding AKE, Table 7 presents summary of reviewed literature. If multiple features and experiments are performed in a study, best results are being reported in following table. *Approach* highlights the underlying key approach used in a study whereas *Features/ Distinctions* highlight the distinctive feature in any study. Some studies have reported evaluation metrics on various numbers of machine-generated keywords. Therefore, to highlight results against various numbers, *K* column in results header denotes the number of keywords extracted. Hence, entries in this column of 5, 10, 15 and 20 denotes results when gold-annotated dataset is compared with machine-generated results, where machine was asked to extract 5, 10, 15 and 20 keywords respectively. *P, R and F* denotes precision, recall an F-measure. It could be noted that, with an increase in keywords, recall tends to increase as chances of match gets higher on the cost of decreased precision.

5. Textual summarization

Automatic TS can be broadly classified into two main classes, namely extractive summarization and abstractive summarization. Extractive summarization refers to the process of summarization where firstly important keyphrases and/or sentences are identified from text. Here, importance of sentence is a function that makes use of multiple sentence features such as lexical, syntactical, statistical and linguistic. Later, these important sentences are concatenated in order to generate a summary. Extractive summarization can also be perceived as classification problem where we have to classify each sentence that whether it should be in summary or not. Primary challenges in extractive summarization include irrelevant information in generated summary due to long sentences, anaphora resolution (sentence in summary might be referring to some concept in preceding sentences) and conflicting sentences. The

Table 7
Literature Summary for AKE.

Study	Dataset	Approach	Features/ Distinctions	Results			
				K	P	R	F
Wu et al. (2005)	500 papers from journals and conferences	Heuristics	Development of Glossary using Human Annotations	5	27.0	31.0	–
				10	19.0	44.0	
				15	15.0	50.0	
				20	12.0	54.0	
Kumar and Srinathan (2008)	Full text articles	Pattern Matching	Compression algorithm LZ78	5	28.8	24.0	–
				10	20.0	32.4	
				15	14.9	36.5	
				20	11.9	38.6	
	708 abstract			5	23.6	16.8	–
				10	15.9	22.5	
				15	11.7	24.9	
				20	08.9	25.3	
Pay (2016)	Inspec	Heuristics	Location, word frequency and	–	44.3	46.9	45.6
	ACM CS		degree Dynamic Threshold	–	28.9	32.8	30.7
Pay and Lucci (2017)	Inspec	Unsupervised	Mean Threshold	–	46.7	50.9	48.7
		Ensemble	Median Threshold	–	42.1	55.9	48.0
Witten et al. (1999)	CTRS	Naïve-Bayes	TF-IDF	Average Match			
				5	0.93		
				10	1.39		
				15	1.68		
				20	1.88		
Medelyan and Witten (2006)	UN AFO	Naïve-Bayes	Domain Specific Thesaurus	–	28.3	26.1	25.2
Hulth (2003)	Inspec	Bagging	Linguistic features	–	25.2	51.7	33.9
Lopez and Romary (2010b)	Sem-Eval 2010	DTC	GROBID, GRISP	5	39.0	13.3	19.8
				10	32.0	21.8	25.9
				15	27.2	27.8	27.5
Huang et al. (2006)	eBooks	SVM	Syntactic and Statistical Features	5	25.0	–	–
				10	15.0		
	Aliweb and NASA			5	27.5	–	–
				10	25.0		
	FIPS and Journal			5	50.0	–	–
				10	37.5		
Turney (2000)	75 Journal articles[b]	GA	Word and POS Features	5	29.0	–	–
				15	17.7		
Berend and Farkas (2010)	Sem-Eval 2010	NB	Phrase, Document, Knowledge Base	–	–	–	22.1
Haddoud and Abdeddaim (2014)	Sem-Eval 2010	Logistic Regression	DPM-Index, Statistical Features	5	44.8	15.3	22.8
				10	36.2	24.7	29.4
				15	28.3	28.9	28.6
Yang et al. (2017)	MAS	CRF	Twenty-Two Features, Mutual Information	–	–	–	70.5
Thomas et al. (2016)	INND	HMM for POS-tagging	Word Count, Statistical Formulation based on Count and POS	–	61.0	57.0	59.0
Xie et al. (2017)	Sem-Eval 2010	Naïve-Bayes	Sequence Mining, TF-IDF	–	–	–	19.8
Mihalcea and Tarau (2004)	Inspec	Graph-based	POS Features	–	–	–	36.2
Rose et al. (2010)	Inspec	Graph-based	Stoplist generation, Words adjoining	–	33.7	41.5	37.2
Wan and Xiao (2008)	DUC 2001	Graph-based	Neighborhood Knowledge	5	26.4	32.5	29.1
				10	28.8	35.4	31.7
				15	28.6	35.2	31.6
El-Beltagy and Rafea (2009)	English[c]	Hybrid	General Features	7	24.1	20.5	18.6
				15			
				20			
	Arabic Wikipedia			–	–	–	19.6
Sterckx et al. (2018)	Online News	Gradient Boosted	Dutch Wikipedia, Contextual	5	56.4	–	31.8
	Online Sports	Decision Trees	information, TF-IDF, Topic		56.2	–	36.7
	Magazines		Modeling		46.4	–	26.8
	Printing Press				47.5*	–	26.5*
Liu et al. (2010)	DUC 2001	Graph-based	LDA for topic modeling	10	28.2	34.8	31.2
	Inspec	(PageRank)		5	35.4	18.3	24.2
You et al. (2013)	Sem-Eval 2010	Hybrid	Position, Statistical	–	26.2	26.8	26.0
Park et al. (2014)	Blogs	LDA	Richness Measure	Self-proposed evaluation metric			

(*continued on next page*)

Table 7 (*continued*)

Study	Dataset	Approach	Features/ Distinctions	Results			
				K	P	R	F
Hong et al. (2014)	One article (Adamic, Zhang, Bakshy, & Ackerman, 2008)	Hybrid	Crowdsourcing	–	62.0	62.0	62.0
Song et al. (2017)	Self-annotated Twitter dataset	Graph-based	Temporal History	–	–	–	49.8
Lahiri et al. (2017)	Email	KNN	Word, Phrase	–	31.9	50.0	38.9
Zhang et al. (2017)	WWW	Graph-based approach	Prior knowledge, Word Graph structure	–	14.8	15.0	14.9
	KDD			–	16.1	16.0	16.0
Beliga (2014)	Croatian News	Graph based	Selectivity	–	–	–	25.9
Alrehamy and Walker (2018)	Inspec	Hybrid	WordNet, POS-tagging, Chunking, WSD	–	40.1	74.2	52.0
	DUC 2001			–	36.4	69.2	47.7
Yang et al. (2018)	20News	LDA	Initial seed queries, PMI-IR	–	83.0	–	–
	OMD			–	81.0	–	–
Ying et al. (2017)	Inspec	Graph-based	Variety of word graphs	–	43.0	40.2	39.6
	900 Articles			–	48.7	49.8	47.8
Jo (2003)	News	MLP	TF-IDF, Word Presence	Not clearly expressed			
Wang et al. (2006)	250 English text	MLP	TF-IDF, Structural Features	–	–	–	31.6
	50 Chinese text			–	–	–	37.5
Sarkar et al. (2012)	210 Journal articles	MLP	Phrase level Features	5	34.0	36.0	–
				10	23.0	48.0	
				15	17.0	53.0	
Aquino et al. (2014)	WICC-43	Auto-encoder	TF-IDF, Word Presence	–	36.2	34.3	34.4
	WICC-166			–	26.6	24.9	25.6
	CAICC-72			–	13.8	17.7	15.3
Jo and Lee (2015)	Inspec	DBN(MLP)	Latent keyphrase	–	–	–	10.8
Zhang et al. (2016)	Twitter data	RNN	Embeddings	–	80.7	81.2	80.9
Marujo et al. (2015)	1827 Tweets	Hybrid	Word Vectors	–	72.0	75.2	73.6
Li et al. (2017)	4272 Chinese abstracts	Graph-based (TextRank)	HMM-based segmentation	–	48.6	49.4	48.9
Alzaidy et al. (2019)	ACM KDD	Bi-LSTM-CRF	Sequence labeling formulation for AKE, Glove 100-d embedding	–	57.8	31.8	41.1
	ACM WWW			–	64.3	28.4	39.4
	KP20K			–	64.2	24.7	35.6
Ray Chowdhury et al. (2019)	General Tweets	Bi-LSTM	Contextualized word embedding (Elmo) and Twitter Glove, character level CNN embedding	–	–	–	85.1
	Disaster Tweets			–	–	–	66.4

[a] These scores are achieved when background knowledge using Wikipedia is not incorporated. For details, please refer to Sterckx et al. (2018).
[b] Study has experimented on total of five data genres. Here, only journal is reported. For details follow Turney (2000).
[c] Study carries extensive experimentation using all datasets reported in Turney (2000), here average results are reported.

aforementioned example in Section 2 is the case of extractive summarization.

Abstractive summarization, on other hand, makes use of natural language processing in order to rephrase the generated summary to form more coherent text. This method requires in-depth document analysis to better identify the underlying concepts residing in document for effective summarization. Hence, linguistic methods are required for thorough analysis and interpretation of text. After identifying key-insight of a document, eventual focus of abstractive summarization is to produce a grammatically sound and coherent summary. This task requires advanced techniques in domain of language generation and modeling; and also serves a major challenge. Hence, amongst these two approaches, abstractive summarization is more challenging problem than that of extractive summarization.

Summarization can also be applied on a single document or a set of documents and is termed as single-document summarization and multi-document summarization respectively. In literature, majority studies focus on the task of extractive summarization by means of keyword extraction and sentence selection. Recently, neural network based approaches are being applied to perform abstractive summarization. Following section firstly describe existing survey studies for summarization followed by state-of-the-art approaches that are compiled in the light of existing surveys as well as by performing literature review.

5.1. *Existing surveys*

There exist multiple surveys regarding extractive summarization (Das & Martins, 2007; Gambhir & Gupta, 2017; Moratanch & Chitrakala, 2017). These existing surveys classify major approaches to extractive summarization into frequency based, machine-learning based, graph based, fuzzy-logic based, query based, semantic analysis and cluster based. This section tends to provide a bird's eye view of existing survey studies. In addition to that, a survey study carried out in Khan and Salim (2014) presents various approaches for abstractive summarization

Review study carried out in Das and Martins (2007) has its major focus set on extractive summarization in single and multiple documents. In the light of this study, major techniques for single-document extractive summarization include Bayesian classifiers, decision tree classifiers, HMM, log-linear models and Neural Networks. On the other hand, approaches for multi-document summarization include template driven approaches, information fusion, maximum marginal relevance, graph based techniques and clustering techniques. Other various forms of summaries are also described in this study that includes short summaries that are being generated by means of learning a translation model between document and its summary. In addition to that, automatic evaluation measures for summarization are also explained including ROUGE, Recall and N-gram matching score. Amongst information theoretic measures, Jensen-Shannon divergence (JS-divergence) and Kullback–Leibler divergence (KL-divergence) are also being used where, JS-divergence being bounded and symmetric measure, is a better evaluation measure than that of KL-divergence.

A review study for abstractive summarization is presented in Khan and Salim (2014) that tends to classify approaches in structure based approaches and semantic based approaches. Structure based approaches capture the important information in the form of cognitive schemas such as templates, trees, ontologies etc. Structure based approaches are further classified in tree based, template based, ontology based, lead-and-body-phrase and rule based methods based on the underlying schema used. On the other hand, semantic based approaches make use of semantic information in order to generate text. In semantic based approaches, verb and noun phrases are usually identified by means of linguistic data. Semantic based approaches are being further classified on the basis of semantic model construction. Major approaches to construct semantic model include multi-modal semantic model, information item based method and semantic graph based methods. Study concludes that majority of the employed approaches for abstractive summarization produce short and cohesive summaries.

Review study presented in Moratanch and Chitrakala (2017) is focused on extractive text summarization. It majorly covers word related and sentence related features that are being used to select the candidate sentences for summary. This review study majorly classifies approaches to extractive summarization in supervised and unsupervised. Supervised approaches are further classified into Bayesian approaches; neural networks based approaches and CRF. Unsupervised approaches are classified into fuzzy-logic based, concept based, graph based and semantic analysis based approaches. It further includes evaluation approaches for summarization that include ROUGE, Precision, Recall and f-score. Survey study also argue about need of devising improved evaluation mechanisms for summarization.

The survey study presented in Gambhir and Gupta (2017) lists the various grounds for classifying techniques of summarization. These grounds include nature of summary generation (extractive/abstractive), nature of input (single/multi-document), output style (indicative/informative), language (multi-lingual/mono-lingual/cross-lingual) and source of input (web/email). Major focus in this review study is on extractive and abstractive summarization. Extractive summarization is further classified based on various techniques including statistical approaches, topic based approaches, graph based approaches, discourse based approaches and machine learning approaches. Further, automatic approaches for summarization are elaborated in a brief fashion along with their comparison with each other. Abstractive summarization, on the other hand, is addressed quite less in the literature in comparison to extractive summarization. This study compiles summaries of all research studies for abstractive summarization in tabular form. Similar tabular approach is followed in order to consolidate state-of-the-art research studies in multi-lingual text summarization. In addition to that, various evaluating schemes for summarization are explored including extrinsic evaluation and intrinsic evaluation. Extrinsic evaluation can be performed by means of reading comprehension and relative assessment. Intrinsic evaluation, on the other hand, can be performed using two measures either quality or informativeness. Further, summarization quality is assessed by means of grammatical analysis, redundancy check, text structure and its coherence, whereas informativeness is evaluated by means of various statistical measures including Relative Utility, ROUGE, Precision, Recall, F-score, and Pyramids etc. In addition to that, comparative analysis is performed for various extractive summarization techniques using DUC dataset. This analysis shows that a clustering based technique for sentence ranking outperform other all techniques for summarization. The authors conclude that feature selection is important to improve summarization. Moreover, this study also put great emphasis on improvement of evaluation measures as pointed out in earlier studies (Moratanch & Chitrakala, 2017) as well. Hence, semantic base evaluation of summaries currently stands as a wide research gap.

In the light of existing surveys, widely used approaches include heuristics, supervised and unsupervised approaches for extractive summarization. Existing survey studies point to relatively less development in abstractive summarization in comparison to extractive summarization. Amongst various surveys, only one study briefly mention employment of RNN to the task of summarization. Therefore, neural network approaches are majorly emphasized in this study. Table 8 highlights the major emphasis and features of existing surveys. In this table PR denotes whether process and repositories to acquire literature is described. FE denotes whether features are discussed that are employed to construct various systems. Lastly, ST denotes whether summary table of reviewed literature is presented in a survey. Value CI denotes comprehensive information, whereas PI denotes partial information.

Table 8
Summary of existing TS surveys.

Study	Major focus/ Distinctions	Coverage	PR	FE	ST
Das and Martins (2007)	Single and multi-document summarization	1998–2010	–	CI	PI
Khan and Salim (2014)	Abstractive summarization, structure and semantic based methods	1999–2012	–	–	PI
Moratanch and Chitrakala (2017)	Extractive Summarization	1998–2015	–	CI	CI
Gambhir and Gupta (2017)	General	1998–2015	–	CI	CI
This study	Neural network approaches, underlying datasets	1998–2019	CI	CI	CI

5.2. Rule/heuristic based approaches

The research carried out in Mittal, Kantrowitz, Goldstein and Carbonell (1999) compiles major heuristics that can be used in order to perform automatic text summarization. It firstly compiles various textual features by means of analyzing newswire articles. These features can be broadly classified into three categories: word or phrase level features, sentence level features and document level features. Later, a sentence ranking mechanism is proposed on the basis of weighted combination of various features that depict linguistic information. The study argues that nature of text is extremely important and it affects the overall results. In the light of this argument, the complexity of respective dataset is taken into account while evaluating the summary results. Experiments show that incorporation of various features affects the overall results. Hence, results are sensitive to the input features.

A heuristic based approach presented in Dalal and Zaveri (2011) performs the tasks of automatic summarization in three phases. First phase is topic-signature generation which is made by means of title words, cue-words, user's specified keywords along with their synonyms. Second phase is sentence scoring and extraction phase, where sentences are scored based on presence of signature terms and sentences are normalized with respect to their lengths. Later, sentences having scores greater than a defined threshold are extracted. Final phase is evaluation, where extracted sentences are evaluated by means of multiple soft-metrics that include topic-coverage, novelty and information content. Study concludes that in order to improve the process of summarization, human way of processing information by means of ML approaches should be used.

Another rule based approach is proposed in Abujar et al. (2017) to summarize Bengali documents. Firstly word related features and sentence-related features are extracted. Later, graph-scoring is employed in order to analyze the relationship between sentence and words. A similar rule based system for Hindi documents is also presented in Gupta and Garg (2016).

Another approach that relies on nine heuristics to perform extractive summarization is presented in Wang, Li, Wang, and Zheng (2017). These heuristics include removal of redundant sentences, various scoring methods to determine the suitability of sentence and greedy deletion algorithm to delete worst sentence from set of summary sentence. Results are evaluated using various document resources that include conference, journal and news articles, legal cases and patents. Experiments are conducted using both words and sentences as a basic unit of summary. In case of words, n-gram combination technique is used to generate meaningful phrases. Proposed method is evaluated using various document resources that include conference, journal and news articles, legal cases and patents. In addition, experiments are also conducted on DUC 2002 dataset for varied length abstracts. For the sake of brevity, results against major classes are presented in Table 9.

5.3. Supervised approaches

Approach proposed in Fattah and Ren (2008) makes use of features in order to employ genetic algorithms (GA) and mathematical regression (MR) for the task of text summarization. Study uses total of ten potential features for training models that include position, positive and negative keywords in sentence and sentence centrality. Total of 150 religious English documents were used, where 50 were used for training i.e. identifying the features and train the model, and rest were used for testing. Precision is used to evaluate the results and experiments conducted on various compression rates of 10%, 20%, 30%. Respective precision values for these compression rates are 43.76, 44.35, 44.94 and 43.12, 43.53, 43.82 for GA and MR algorithms respectively.

A semi-supervised approach using SVM and Naïve–Bayes is proposed in Wong et al. (2008) which aims to reduce labeling time as supervised approaches require labeled data. Various sentence-related features are employed that are classified as relevance, event, content and surface features. Relevance features evaluate a sentence based on its level of relatedness with other sentences. Event features represent sentences on the basis of events that exist in sentences. Content features measure a sentence by means of content conveying words. Whereas, surface features makes use of extrinsic aspects of a sentence. These various features are combined and results show that the combination improves summarization performance significantly. Primary challenge in supervised approaches is to develop annotated data, that requires a great deal of time. Thus, to reduce the time required for data labeling, co-training is employed using SVM and Naïve-Bayes. The basic intuition regarding co-training is to makes use of both labeled and unlabeled data, which can be regarded as a semi-supervised learning approach. Experiments are conducted using co-training approach as well. Results show that co-training approach is effecient as it saves considerable amount of labeling time and it also achieves comparable performance in comparison to its supervised counterpart.

Mathematical model of Statistics (MMS), HMM, CRF and Gaussian Mixture Models (GMM) are employed in Yu and Ren (2009) for cross-language text summarization in Chinese and Japanese. To train the models, several features are employed including sentence position, sentence centrality, numeric characters, similarity of sentence with title. Around 200 Chinese articles from web pages and 120 Japanese's articles from NTCIR corpus are selected for evaluation of the algorithms. Precision is used as evaluation measure and it is calculated for compression rates of 10%, 20% and 30%. Experiments are conducted in order to check the impact of various features on overall precision when MMS are employed. Results show that GMM tends to outperform other approaches in all cases against both languages.

5.4. Unsupervised approaches

Unsupervised approaches employ well-known constructs to generate intermediate results. Later, acquired results are processed using heuristics or algorithms to finalize the insights. A pioneer study to employ lexical chains for the task of summarization is presented in Barzilay and Elhadad (1999). Lexical chain construction consists of selection of candidate words followed by construction of chain using each candidate word in an iterative fashion. Therefore firstly, Barzilay and Elhadad (1999) discusses an

Table 9
Literature summary for TS.

Study	Dataset details	Approach	Features/ Distinctions	Results C	P	R	F	R1	R2	RL
Mittal et al. (1999)	Synthetic Dataset	Heuristics	Word, Sentence, Document	Effect of various features and heuristics measured						
Dalal and Zaveri (2011)	80 Undergraduate Assignments	Heuristics	Keywords, Title Words, Position	Qualitative results						
Abujar et al. (2017)	Three Bengali Documents	Heuristics	Word, Sentence	Not clearly stated						
Gupta and Garg (2016)	30 Hindi Documents	Heuristics	External Lexicons	Accuracy: 96						
Wang et al. (2017)	Conference	Heuristics	N-Gram Based Word Processing, Greedy Approach to Delete Sentences, Various Scoring Mechanisms	–	–	–	–	56.1	31.3	–
	Journal			–	–	–	–	62.5	38.6	–
	Legal Cases			–	–	–	–	43.6	22.8	–
	DUC 2002			–	–	–	–	62.0	41.0	–
	Patent			–	–	–	–	74.6	54.5	–
	DUC 2002[50]			–	–	–	–	41.3	18.4	–
	DUC 2002[100]			–	–	–	–	47.4	22.7	–
	DUC 2002 [200]			–	–	–	–	52.3	27.7	–
Fattah and Ren (2008)	150 Religious English Scripts	GA	Position and Words' Sentiment in Sentence	10	43.7	–	–	–	–	–
				20	44.4					
				30	44.9					
Wong et al. (2008)	Duc 2001	Semi-Supervised	Co-Training of SVM and NB, Multiple Features	–	–	–	–	39.6	11.6	35.8
Yu and Ren (2009)	200 Chinese Web Pages	GMM	Named Entity Counts, Position and Features	10	60.5	–	–	–	–	–
				20	62.1					
				30	61.3					
	120 Japanese Articles NTCIR			10	61.9	–	–	–	–	–
				20	63.0					
				30	62.5					
Barzilay and Elhadad (1999)	30 Magazines	Lexical Chains	Wordnet, Heuristics	Empirical verification of results						
Silber and McCoy (2000)	Any Text or HTML	Lexical Chains	Wordnet	–	–	–	–	–	–	–
Reeve et al. (2006)	Oncology Clinical Trial Texts	Lexical Chains	Domain-Specific Pruning	–	90.0	92.0	–	–	–	–
Lynn et al. (2018)	600 News Articles	Lexical Chains	Wordnet	–	–	–	–	59.8	41.9	–
García-Hernández and Ledeneva (2009)	DUC 2002	K-means Clustering	TF-IDF	–	44.8	44.9	44.6	–	–	–
Sarkar (2012)	38 Bengali Documents	TF-IDF	Sentence Position Index	–	36.6	50.6	41.7	–	–	–
Wu et al. (2015)	100 Single Documents	Spectral Clustering	LexRank	10	61.1	43.1	48.7	–	–	–
				20	51.7	40.8	44.9			
Mihalcea (2004)	DUC 2002	Graph based	TextRank	–	–	–	–	50.3	–	–
Litvak and Last (2008)	DUC 2002	Graph based	Hits Algorithm	–	–	–	39.2	–	–	–
Ganesan et al. (2010)	Opinosis	Graph based	Surface order of words					32.7	10.7	
Gong and Liu (2001)	CNN Worldview News Program	LSA	Various Weighting Schemes	–	53.0	61.0	57.0	–	–	–
Steinberger and Jezek (2004)	Reuters Data Having 20+ Sentence	LSA	Various Similarity Measures	Average Cosine Similarity 77.15						
Ozsoy et al. (2011)	Duc: 2002	LSA	Various Sentence Selection Approaches	–	–	–	23.0	–	–	–
	Duc: 2004			–	–	–	10.2	–	–	–
	Turkish			–	–	–	32.0	–	–	–
Lin and Bilmes (2011)	DUC 2004	Submodular function	Diversity Rewarding Function	–	–	–	–	39.3	–	–
	DUC 2005			–	–	–	–	–	08.4	–
	DUC 2006			–	–	–	–	–	09.8	–
	DUC 2007			–	–	–	–	–	12.3	–
Fang et al. (2017)	10 Chinese News Articles	Graph-based	Sentence-Sentence and Word-Sentence Relations For Scoring	–	–	–	59.4	69.7	60.6	66.1
	DUC 2002			–	–	–	52.4	52.6	25.8	45.1
De la Peña Sarracén & Rosso, 2018)	DUC 2002	Graph-based	Betweeness Centrality	BLEU: 55.81	38.3	–	–			
Kaikhah, 2004)	85 News Articles	MLP	Drop-Out	Accuracy: 99.0						
Kågebäck et al. (2014)	Opinosis	MLP	Embeddings	–	–	–	–	57.9	22.9	29.5
Yousefi-Azar & Hamey (2017)	SKE	MLP (AE)	Local Vocabulary, Noisy Encoders, TF-IDF, Various Noise Distributions	–	16.8	23.2	19.0	62.0	–	–
	BC3			–	–	–	–	12.0	–	–

(continued on next page)

Table 9 (*continued*)

Study	Dataset details	Approach	Features/ Distinctions	Results						
				C	P	R	F	R1	R2	RL
Tsai et al. (2016)	MATBN	CNN	Lexical Features	–	–	–	–	52.9	43.2	48.4
Cao et al. (2016)	Duc 2005	CNN	Greedy Algorithm for Sentence Selection	–	–	–	–	37.0	06.9	–
	Duc 2006			–	–	–	–	40.9	09.4	–
	Duc 2007			–	–	–	–	03.9	11.6	–
Wang et al. (2018)	Gigaword	CNN	Multi-Attention, LDA Topic Modeling, RL	–	–	–	–	36.9	18.3	34.6
	Duc 2004			–	–	–	–	31.2	10.8	27.7
	LCSTS			–	–	–	–	39.9	21.6	37.9
Rush et al. (2015)	Duc 2004	RNN	Viterbi Decoding	–	–	–	–	28.2	08.5	23.8
	Gigaword			–	–	–	–	31.0	12.6	28.3
Nallapati et al. (2016)	CNN/DM	RNN ED	Embeddings, Linguistic Features, Temporal and Hierarchal Attention	–	–	–	–	35.5	13.3	32.6
Paulus et al. (2017)	CNN/DM	RNN ED	Intra-Attention, RL	–	–	–	–	41.2	15.8	39.1
	NYT			–	–	–	–	47.2	30.5	43.3
Ma et al. (2018)	Movie & TV	LSTM-based joint model	Multi-View Attention, Joint Model for Summarization and Sentiment Mining	–	–	–	–	14.5	04.8	13.4
	Sports			–	–	–	–	17.8	04.8	17.6
	Toys & Game			–	–	–	–	18.4	5.0	17.7
Chen et al. (2015)	MATBN	RNN	Language Model	10	–	–	–	53.3	43.9	48.3
				20				58.0	47.8	52.2
				30				63.9	54.0	57.4
				25[b]				42.3	28.1	36.2
Nallapati et al. (2016)	CNN/DM	RNN	Binary Sentence Classification	–	–	–	–	26.2	10.8	14.4
Collins et al. (2017)	ScienceDirect Open-Access Articles	LSTM	AbstractROUGE, Ensemble Network	–	–	–	–	–	–	32.0
Khatri et al. (2018) [a]	Ebay Reviews	RNN, CNN	Document Context Vector, Attention	Ab	–	–	–	33.0	31.0	38.0
				Ex	–	–	–	39.0	37.0	30.0
Rekabdar et al. (2019)	CNN/DM	GAN	LSTM ED & CNN as Gen. & Disc., Policy Gradient, Attention	–	–	–	–	37.9	15.7	39.2
Song et al. (2019)	CNN/DM	CNN and LSTM ED	Improved Phrase Extraction	–	–	–	–	34.9	17.8	–

[a] Ab and Ex in CR column in this row represents best results against abstractive and extractive summarization respectively.

[b] This compression rate is used to evaluate generated result with human-written abstractive summaries.

improved way of constructing lexical chain using shallow parsing. After lexical chain constructions, various heuristics and measures are employed to score these constructions in the light of empirical verification. These heuristics include length and homogeneity index that requires number of distinct occurrences to compute. Using this information, lexical chains are scored. Furthermore, after scoring, three heuristics are proposed to generate the summary. The proposed approach does not offer any way to control the summary length and deems anaphora resolution in the eventual summary as one of the challenge. Results are empirically validate against thirty magazine texts. Results show that proposed approach provides reasonable results. Other studies employing lexical chains for AKE include Lynn et al. (2018), Reeve, Han, and Brooks (2006) and Silber & McCoy (2000).

A TF-IDF based approach that employs term selection and weighting for summarization is proposed in García-Hernández and Ledeneva (2009) which further uses clustering algorithm to generate summary. Similar approach along with sentence feature is used in Sarkar (2012) in order to generate non-redundant summary. Another study employing TF-IDF in MapReduce framework to perform AKE is presented in Mackey and Cuevas (2018).

Spectral clustering based approach to summarization is also introduced in Wu, Shi, and Pan (2015). When clustering is performed, each cluster carries data points that are similar to each other in one way or the other. Thus by the very virtue of clustering concept, it is best suited for documents that carry multiple themes and/or address various entities and aspects. In the context of summarization; If a cluster represents a theme, then various clusters can represent various entities/aspects of a document, where document can be represented using TF-IDF. Thus, this methodology provides summary that is relevant to the document topic or query.

Fuzzy logic based systems are also employed to perform summarization by using rules and features defined in the form of knowledge base (Patil & Kulkarni, 2014). These systems are simple and quite flexible in nature as they can handle noisy data as well. A multi-document summarization system using fuzzy logic is also proposed in Tsoumou, Lai, Yang, and Varus (2016).

Graph based approaches (Litvak & Last, 2008; Malliaros & Skianis, 2015; Mihalcea, 2004; Sevilla, Fernández-Isabel, & Díaz, 2016), aids in identifying major themes/topics covered in the document. In order to perform summarization, every sentence is considered as graph node and weighted edges represent the strength of similarity between sentences. Later sentences are ranked, based on various nodes ranking mechanism. Major advantage of these approaches is that they do not require any linguistic/domain-specific information, which makes these approaches domain as well as language independent.

Latent Semantic Analysis (LSA) is also employed in many studies (Gong & Liu, 2001; Ozsoy et al., 2011; Steinberger & Jezek,

2004) in order to extract meaning of sentence or to perform sentence ranking, which are later used in order to generate summary. These methods cannot exploit information of word order and syntactic relations as they solely rely on information in input text.

Another unsupervised approach employing submodular function is presented in Lin and Bilmes (2011). Each function makes use of two major terms. One term deals with correctness and relevance of generated summary whereas other term is focused on making summary more diverse. This study argues that ROUGE evaluation metrics itself is closely related to submodular functions. Hence, application of submodular approach results into improved ROUGE scores. Results are reported on various DUC datasets with ROUGE-1 of 39.3 on DUC 2004. Whereas ROUGE-2 of 8.38, 9.75 and 12.33 are achieved on DUC 2005, 2006 and 2007 respectively. Other unsupervised approaches include (De la Peña Sarracén & Rosso, 2018; Fang et al., 2017).

5.5. Neural network approaches

After the advent of word-vectors, there is significant increase in employment of neural-network frameworks to perform TS. Initial neural nets studies often relied on manually crafted features. Whereas, current studies rely on word-vectors. Amongst various deep-learning frameworks, RNN-based LSTMs are among the widely used techniques. Following section briefly captures advancements made in the light of widely used neural network frameworks.

5.5.1. Multi-layer perceptron

A research study to employ neural networks for summarization is proposed in Kaikhah (2004). This study trains the neural network based on the sentence presence in final summary. Thus, model learns the inherent features of sentences that are part of final summary by means of learning weights. Weights learnt after model training are later pruned, based on the threshold, i.e. if weight value is smaller than the threshold, that weight is reduced to zero. Similarly, if a hidden neuron no longer carries information to communicate further because of zero weights, such neurons are also dropped out. This process helps in order to remove uncommon features and to provide generalization up to some extent. After handling uncommon features, next step involves combining the effects of prominent features that are identified via model learning earlier. This is carried out by performing adaptive clustering on hidden layer activation values. After applying clustering, each cluster can be identified by means of its centroid and frequency. Thus, through this feature fusion processing, sentences are ranked. Proposed approach for summarization is applied on a three-layered feed forward neural network in this study to perform extractive summarization. Results are reported on self-made dataset of 85 news articles carrying total of 2835 sentences. Sentences were manually marked by humans and later accuracy is calculated by total number of true classifications. Experiments were performed using real-valued feature vectors, discretized real-valued into intervals and discretized real-valued into single value which resulted on the accuracy of 93, 96 and 99.

A pioneer study to use word vector representation for summarization is proposed in Kågebäck et al. (2014) . This approach uses word embedding in order to calculate sentence similarity. There are multiple ways to calculate phrase level similarity including word vectors addition or unfolding recursive auto-encoder that tends to use binary parse tree in order to generate phrase embeddings. The proposed similarity approach is applied on dataset presented in Ganesan et al. (2010). Proposed approach resulted in optimal scores of 57.86, 22.96 and 29.50 using ROUGE-1, ROUGE-2 and ROUGE-4 respectively.

Another approach employing autoencoders to perform summarization on emails is presented in Yousefi-Azar and Hamey (2017). It incorporates random noise to input while learning features using autoencoders. In addition, the process employs multiple such autoencoders generating ensemble of noisy autoencoders (ENAE). Also, various features that include TF-IDF, term frequency learned using local vocabularies, and Gaussian and uniform distributions for noise incorporation are used. Comprehensive comparative analysis is performed using all these features. Results show that term frequency learned using local vocabularies along with noisy auto encoders improve results.

5.5.2. Convolution neural networks

CNNs are being used in conjunction with Multi-layer perceptron in Tsai, Hung, Chen, and Chen (2016) to perform extractive summarization. Two different CNNs (CNN1 and CNN2) are proposed where each of them uses set of fifty filters and each filter has a context size of five consecutive words. Moreover, max pooling is used in both CNNs. Both of these architectures outperform state-of-the-art approaches on MATBN corpus (Wang, Chen, Kuo, & Cheng, 2005) with ROUGE-1/ROUGE-2/ROUGE-L measures of 50, 40.7, 46 and 52.9, 43.2, 48.4 for CNN1 and CNN2 respectively. Amongst these two, CNN2 outperform CNN1 as CNN2 makes use of extra indicative features as well.

Another approach for extractive summarization on multiple documents is presented in Cao et al. (2016) . This particular approach is focused on summary generation based on user-query. The major contribution of this approach is that it models query relevancy as well as sentence ranking in a joint fashion. To achieve this, resultant process consist of three major layers. First layer applies CNN to generate the embeddings against query as well as sentences. This CNN layer is followed by pooling layer. In the pooling layer, a weighted sum pooling operation is used, that is performed over sentence embeddings generated via CNN in earlier phase to represent the clusters of documents. This polling layer is followed by a sentence ranking layer that uses cosine similarity to calculate and rank sentence with that of document cluster generated via pooling layer. Lastly to select summary sentences, proposed approach makes use of greedy algorithm. Results are reported on DUC corpora from 2005, 2006 and 2007. Proposed approach resulted in ROUGE-1 and ROUGE-2 scores of (37.01, 6.99), (40.90, 9.40) and (43.92, 11.55) on DUC corpus of 2005, 2006 and 2007 respectively. This approach tends to model the query ranking and sentence ranking in a joint fashion. In addition to that, an attention mechanism that simulates the human attention behavior for query focused summarization is incorporated.

An approach for abstractive summarization using CNN is presented in Wang et al. (2018). Its major distinction is incorporation of topic information to improve the summarization process. Encoder and decoder are units comprising stacked convolutions. To incorporate topic modeling information, topic embeddings are prepared using LDA model. Hence, these embeddings result into bias generation that consequently results in improved summarization results. Furthermore, reinforcement learning (RL) is applied using teacher forcing algorithm (Williams & Zipser, 1989). The proposed approach is evaluated on three datasets including Gigaword corpus, DUC 2004 and Chinese short text documents (LCSTS). Results show that proposed approach outperforms state-of-the-art deep learning solutions for abstractive summarization.

5.5.3. Recurrent neural networks

Abstractive Summarization: The research work presented in Rush, Chopra, and Weston (2015) uses multi-layered neural network language models with various encoders (bag-of-words, convolution and attention based). Finally Viterbi decoding is used to generate abstractive summaries. Results are evaluated on DUC 2003 and 2004 datasets. Experiments show ROUGE-1/ROUGE-2/ROUGE-L measure of 28.18, 8.49, 23.81 on DUC 2004 dataset and respective measure on GigaWord dataset are 31.00, 12.65 28.34, where best human evaluator ROUGE-1 score is 31.7. The major contribution of this study is the use of neural probabilistic model with generation model to create summaries. An extension of this study is presented in Chopra, Auli, Rush, and Harvard (2016) .

The study proposed in Nallapati, Zhou et al. (2016) makes use of attention based encoder-decoder RNN that was initially proposed for machine translation in Bahdanau, Cho, and Bengio (2014). Encoder consists of a bi-direction GRU whereas decoder is simple GRU. Word embeddings are used as input to the system. Encoders also employ linguistic features which are not used in abstractive summarization before. In addition to that, attention mechanism is employed at both word-level and sentence-level in a joint fashion; that forms hierarchical attention based models. These models tend to repeat phrases in resultant summary. Hence, temporal attention models (Sankaran, Mi, Al-Onaizan, & Ittycheriah, 2016) are applied to overcome this challenge . On CNN/Daily Mail dataset, best ROUGE-1/ROUGE-2/ROUGE-L measures of 35.46, 13.30, 32.65 are achieved using temporal attention based model, which outperformed hierarchical attention based model as well as baseline model without attention. For DUC dataset, proposed model outperforms existing models in terms of ROUGE-2 and ROUGE-L measures.

Another encoder-decoder RNN based model is proposed in Paulus et al. (2017) which combines RL with standard supervised methods to perform abstractive summarization. It proposes an RNN based encoder-decoder to achieve an intra-attention model. It also incorporates maximum likelihood and RL to define a hybrid objective function to train the model. Results are presented on CNN/ Daily Mail Corpus and New York Times datasets. Datasets split percentages are 95, 5 and 5 for training, development and testing dataset respectively. On CNN/Daily Mail dataset, experiments show ROUGE-1, ROUGE-2 and ROUGE-L scores of 41.16, 15.75 and 39.08 with RL along with intra-attention whereas scores are 39.87, 15.82, and 36.90 with intra-attention model for hybrid training function (including both RL and maximum likelihood). On New York Times dataset, experiments show ROUGE-1, ROUGE-2 and ROUGE-L scores of 47.22, 30.51 and 43.27 with RL with no intra-attention whereas scores are 47.03, 30.72 and 43.10 for hybrid training function (including both RL and maximum likelihood) with no intra-attention model.

A rather recent approach to jointly train machine learning models for text summarization and sentiment classification is presented in Ma, Sun, Lin, and Ren (2018). In this work, authors have treated sentiment classification as a special case of text summarization, hence, presenting a joint model for these two problems. Proposed model consists of three major components that include an input encoder, a summary decoder followed by sentiment extractor. Text encoder consists of BI-LSTM units to provide more meaningful representation of text to subsequent model layers. Summary decoder consists of uni-directional LSTM units along with multi-view attention to deal with the joint model as well as a word generator. After decoding the text, last layer firstly combines all summary sentiment vectors and input and it later performs max-pooling to find the optimal sentiment context vector. This is finally passed to an MLP unit with RELU to determine the sentiment label, where five-class sentiment classification is being performed. Experiments are conducted on Amazon SNAP Review Dataset (SNAP). Three benchmarks are created from SNAP each covering genre of Movies & TV, Sports & Outdoors and Toys & Games respectively. Authors has presented an ablation study to determine the effect of various parameters. Overall results against all three datasets are presented in Table 9. Key thing to note is that underlying dataset consists of reviews, hence, it carries informal text and lots of noise. Nonetheless, proposed approach outperformed state of the art models when applied on both tasks independently.

Extractive Summarization: An approach for extractive summarization based on RNN is proposed in Chen et al. (2015) in an unsupervised fashion. This study employs RNN based-Language-modeling approach to perform text summarization. RNNLM model has the capability to render the relationship among words in long-span. A hierarchical training strategy is devised to train the model. Experiments are performed on Chinese dataset which resulted into ROUGE-1/ROUGE-2/ROUGE-L measures of 53.3, 43.0 and 48.3 respectively. Extensive experimentation is performed in this study with respect to multiple features.

Another approach that uses RNN to perform extractive summarization is presented in Nallapati, Zhai, and Zhou (2016) . This approach treats the problem of summarization as sequential problem, where every sentence is classified via binary classifier representing its involvement in final summary. This approach presents a novel training method that trains model in an abstract way, thus eliminating the need of generating extractive labels. On the CNN/ Daily Mail corpus, proposed system results into ROUGE-1, ROUGE-2 and ROUGE-L scores of 26.2, 10.8 and 14.4 respectively.

A recent study proposed in Collins, Augenstein, and Riedel (2017) makes use of LSTMs, word vectors and ensemble networks to perform summarization on scientific articles. One of the major contributions of this study is the generation of dataset, which includes ScienceDirect publications having author's highlights. The author highlights against any article is considered as its gold standard summary. Major observation while analyzing author's highlights is that most of the time, complete sentences that are residing in the article are made part of highlights as it is. Thus, copy-paste score is used in study against document sections. This score represents how much content within a section appears in article's summary. To improve the process of extractive summarization, when applied on scientific publications, various constructs are applied. Some of these include ROGUE-L application for training data extension and AbstractROUGE. AbstractROUGE learns features of abstract provided in a scientific article. Results show that incorporation of AbstractROUGE improves performance, but system can still provide better results if it is not included. Similarly, experiments are conducted in order to observe the effect of ROUGE-L measure on increasing training data and later performing summarization based on this extended set of data. Results with respect to ROUGE-L show that increase in training data improves the overall generated summary. Experiments are also conducted based on various features incorporation. Results show that ensemble networks outperform the rest in terms of ROUGE measure.

A rather recent study to perform both extractive and abstractive summarization on eBay reviews is carried out in Khatri, Singh, and Parikh (2018) that apply various DL-frameworks including both RNN and CNN. One of the major distinctive feature used in the study with respect to eBay is construction of a context vector. Using seller's information about a product and user's data, firstly, a context vector is created and is being used during training. It performs extensive experimentation with and without context vector. Each experiment is further performed on abstractive and extractive summarization. Best results against both summarization approaches are presented in Table 9.

Another recent study employs Generate Adversarial Networks (GANs) to perform abstractive text summarization is presented in Rekabdar et al. (2019). The proposed model use Bi-LSTM Encoder decoder framework as generator whereas CNNs are used as discriminator model. Monte-Carlo search and Policy Gradient techniques are applied to train the model. Experiments are conducted on CNN corpus and shows promising results. Another approach employing CNN based decoder and LSTM based decoder for abstractive summarization is presented in Song et al. (2019).

5.6. Conclusion

Text Summarization is being widely used across multiple applications. Many search engines display the highlights of document as per the user query. It is of great importance when audience requires a brief and concise overview of a document. For extractive summarization, many researches rely on extraction of keyphrases to check the saliency of sentences later. Some studies perceive it as classification problem where each sentence is classified on the basis of its features to be part of summary or not. As far as abstractive summarization is concerned, deep learning approaches are being extensively used. The attention based mechanisms improve the overall results.

One major challenge in summarization is handling the subjectivity of gold standard summaries and effective way to incorporate this subjectivity during evaluation. This issue becomes very apparent while evaluating abstractive summaries, as they represent similar semantics but with different lexical outlook. Hence, issues in widely used evaluation metric ROUGE, have been studied in past years (Ermakova, 2012; Schluter, 2017). In addition, some researchers have proposed improved metrics to address the shortcomings for ROUGE. A new evaluation metric SERA (Summarization Evaluation by Relevance Analysis) is presented that is focused majorly on scientific articles summarization (Cohan & Goharian, 2016). Others have extended ROUGE metric by incorporation of new concepts such as word embedding (Ng & Abrecht, 2015) or edit-distance (Ermakova, 2012).

The need of better evaluation metrics has been proposed in earlier surveys as well. Now, there exist several metrics that exploit semantics, but the current challenge is their employment during evaluation. These new metrics should be streamlined and adopted so that their potential pros and cons can be identified. Currently, there seem two apparent reasons for not using the new metrics for evaluation. First one is due to lack of awareness. Community might not be aware of alternate evaluation metrics; hence, to mitigate this, new and improved metrics should be made part of evaluation tracks in various conferences and workshops. Other reason lies in performing the comparative analysis of results. Therefore, it is of utmost importance to study and analyze the effect of the new metrics on existing datasets to provide baselines for comparison.

Table 9 presents summary for studies reported in this survey to solve the problem of TS. If multiple features and experiments are performed in a study, best results are currently being reported here. Some studies have evaluated generated summary using various compression rates. Therefore, It harness this information, in following table, in the ***Results*** header, *C* stands for compression rate, and its values are denoted in percentages. Whereas *P, R, F, R1, R2* and *RL* denote Precision, Recall, F-Measure, Rouge-1, Rouge-2 and Rouge-LCS. *Approach* highlights the underlying key approach used in a study whereas *Features/ Distinctions* highlight the distinctive feature in any study. Moreover, in *Approach* column, RL and ED denotes Reinforcement Learning and Encoder-Decoder framework respectively. All evaluation metrics reported are in percentages.

6. Toolkits and online resources

Due to abundant data in modern era, both tasks under study carry great importance. Thus, many researchers have contributed their research efforts in order to develop tools that could aid fellow community while performing these tasks. Some of these tools are generic in nature whereas others focus on a specific problem. Table 10 lists some of the open-source solutions, code repositories and online websites to perform these tasks, along with programming language and active web links,[3] where ever applicable.

Table above only provides brief subsets of available websites and code repositories to carry out the task of AKE and TS. Most of the online summarizers employ extractive summarization. On the other hand, thanks to open-source initiatives, researchers have publicly

[3] All the shared URLs are visited in 2019.

Table 10
Tools available for AKE and TS.

Tool Name	Prog. Language/ Offered Language	Tasks Solved	Available Link
MEAD Summarizer	Perl	Summarization	http://www.summarization.com/mead/
Gensim	Python	Summarization, Keyword Extraction, Topic Modeling	https://github.com/christophfeinauer/kpex
RAKE	Python	Keyword Extraction	N.A.
KEA	Java	Keyword Extraction	http://community.nzdl.org/kea/download.html
Jtopia	Java	Keyword Extraction	https://github.com/srijiths/jtopia
Topia	Python	Keyword Extraction	Included in python
PKE	Python	Keyword Extraction	https://github.com/boudinfl/pke
OneClick Terms	English, Dutch, Chinese and various others	Keyword Extraction	https://terms.sketchengine.eu/
Translated Labs		Keyword Extraction	https://labs.translated.net/terminology-extraction/
FiveFilters		Keyword Extraction	https://fivefilters.org/term-extraction/
maui-indexer	English, French and Spanish	Keyword Extraction	https://code.google.com/archive/p/maui-indexer/
VocabGrabber		Keyword Extraction	https://www.visualthesaurus.com/vocabgrabber/
TextLenAn	C and python English, French, German, Italian and Spanish texts	Keyword Extraction, extractive summarization	http://texlexan.sourceforge.net/
word-fish https://github.com/word-fish/wordfish-python	Python	Keyword Extraction	https://vsoch.github.io/2016/2016-wordfish/
MontyLingua	Java and Python English	Keyword Extraction, Sentence Generator, Lemmatizer, POS Tagger, Chunking	http://alumni.media.mit.edu/~hugo/montylingua/
LingPipe	Java	Keyword Extraction and various other NLP constructs	http://alias-i.com/lingpipe/web/download.html
TextSummarization	English	Summarization	http://textsummarization.net/text-summarizer
Resoomer	French, English, German, Italian, Spanish	Summarization	https://resoomer.com/en/
AutoSummarizer	Egnlish	Extractive Summarization	http://autosummarizer.com/
Tools4Noobs	English	Extractive Summarization	https://www.tools4noobs.com/summarize/
SUMMA https://summanlp.github.io/textrank/	English	Keyword extraction, Extractive Summarization	https://pypi.org/project/summa/
santhoshkolloju	Python	Abstractive Summarization	https://github.com/santhoshkolloju/Abstractive-Summarization-With-Transfer-Learning
ChenRocks	Python	Abstractive Summarization	https://github.com/ChenRocks/fast_abs_rl

shared their code-bases to perform abstractive summarization, mostly in Github. Majority of these code-bases employ Python framework with Keras, PyTorch and Tensorflow being the widely used libraries at the backend.

7. Conclusion

This study majorly emphasizes on consolidating the research work related to automatic keyword extraction and text summarization. As AKE is focused on extracting keywords that highlights the potential information pointers, and TS is dedicated to present key information in a concise and brief fashion. Therefore, due to similarity in nature of their goals, both of these research problems are being reviewed in the course of this study. In the process of reviewing individual existing studies, major focus was on describing the basic idea along with the proposed method to solve a particular problem. The key contribution and the novelty, if any, are also incorporated against respective studies. This survey study tries to present consolidated picture of what has been done so far by means of describing existing survey studies and recent research. Primary distinctions of this survey include consolidations of relevant existing surveys; extensive review of neural network approaches with major emphasis on deep learning frameworks; brief overview of other non-neural network approaches including heuristics, supervised and unsupervised; consolidation of datasets used by reported literature and a brief overview of available web-applications and open-source toolkits to perform AKE and TS.

For AKE, it is concluded that the majority of work is performed using unsupervised approaches. While most recently, some studies have employed deep learning architectures for AKE. As there exist wide variety of neural architectures and parameters in deep architectures. Hence, application of various deep learning architectures to perform AKE and their respective impacts on results are yet to be analyzed. Also, the impact of various variables on overall results as well as hyperparameters tuning, is an open area in this domain.

As far as TS is concerned, majority of the research studies in the past have focused on extractive summarization. Extractive summarization is focused on extracting best representative sentences against any document. Abstractive summarization, on the other hand, refers to automatic generation of summary that is expressed using natural language. Amongst these two, abstractive summarization is more challenging than extractive summarization due to involvement of natural language synthesis. Recently, by the means of deep learning approaches, there is a great deal of progress regarding abstractive summarization.

If the application of deep-learning (DL) is analyzed across these two problems, TS seems to have progressed a lot in comparison to AKE. Primary reasons for TS progress is availability of datasets that are large enough to train DL frameworks. Summarization datasets from DUC tracks and CNN/DM carry huge amount of data, that was manually summarized for effective evaluation of TS algorithms. Due to availability of these huge resources, DL frameworks are being applied and are presenting state-of-the-art results in TS especially abstractive summarization. On the other hand, AKE datasets are not that large in terms of size to be readily available for DL. Recently, researchers have applied crowdsourcing as well as semi-supervised learning to annotate large datasets for AKE. In addition, some have compiled existing datasets together to develop a larger one.

In addition to these approaches, some researchers have employed heuristics to develop datasets for both AKE and TS. These are usually applied while developing resources for scientific articles. In order to construct summary datasets for bulk of scientific articles, author's highlights are being treated as gold-standard summary. Similarly for AKE, author-keywords are used as gold-standard set of keywords. Hence, when such insights are extracted and combined from bulk of articles, resultant datasets for AKE and TS can be employed for DL.

Keeping the application of deep learning approaches and their increasing employment across multiple domains in view, it is also of interest to design techniques that efficiently utilize energy and provide low-cost solutions while maintaining the accuracy of current systems. These type of solutions require understanding of the datasets at hand along with the idea about current and future computational requirements as discussed in Sze, Chen, Yang, and Emer (2017). Such type of trend analysis on the basis of state-of-the-art employment of DNN in TS can serve as an open future direction.

As both problems are highly dependent upon the datasets at hand, hence, nature of datasets plays an important role. Currently, datasets can be classified in two major classes. One that is focused on formal full-length documents having well-formatted sentences. Other one is focused on short text such as emails and tweets, which can carry formal, semi-formal and informal texts. Model generated and trained on one type might not be readily applicable to the other one due to varied sentence structure. Hence, robust unified algorithms are required which can perform well independent of underlying data. One such approach is unsupervised graphs, that rely on various graphs and centrality measures to perform AKE and eventual TS.

Another major open research area is related to evaluation measures. Mostly evaluation measures perform exact matching or partial matching between systems generated results and gold-standard results. There now exist measures which exploit semantic information but their performances are yet to be analyzed. Furthermore, their employment during evaluation is in itself a challenge, as most studies are still reporting existing evaluation metrics for the sake of performance comparison. Hence, it is of immense importance to encourage the employment of new metrics alongside ROUGE in relevant conference tracks and workshops.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

Abilhoa, W. D., & de Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation, 240*(Aug (Suppl C)), 308–325.

Abujar, S., Hasan, M., Comilla, B., Shahin, M., & Hossain, S. A. (2017). A heuristic approach of text summarization for Bengali documentation. *8th international conference on computing, communication and networking (8th ICCCNT), 2017 8th international conference on.* IEEE.

Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and Yahoo answers: Everyone knows something. *Proceedings of the 17th international conference on world wide web* (pp. 665–674). .

Alrehamy, H., & Walker, C. (2018). Exploiting extensible background knowledge for clustering-based automatic keyphrase extraction. *Soft Computing, 22*(Nov (21)), 7041–7057.

Alzaidy, R., Caragea, C., & Giles, C. L. (2019). Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. *The world wide web conference* (pp. 2551–2557).

Graff, David, and Christopher Cieri. English Gigaword LDC2003T05. Web Download. Philadelphia: Linguistic Data Consortium, 2003. Available: https://catalog.ldc.upenn.edu/LDC2003T05 [Accessed 03 June 2019].

Sandhaus, Evan. The New York Times Annotated Corpus LDC2008T19. DVD. Philadelphia: Linguistic Data Consortium, 2008. Available: https://catalog.ldc.upenn.edu/LDC2008T19 [Accessed 03 June 2019.].

Graff, David. The AQUAINT Corpus of English News Text LDC2002T31. Web Download. Philadelphia: Linguistic Data Consortium, 2002. Available: https://catalog.ldc.upenn.edu/LDC2002T31 [Accessed 03 June 2019].

Microsoft, 2019, “Microsoft Academic.” [Online]. Available: http://academic.research.microsoft.com/. [Accessed 3 June 2019].

Wikipedia, 2019, “Wikimedia Downloads.” [Online]. Available: https://dumps.wikimedia.org/ [Accessed 3 June 2019.].

Aquino, G. O., Hasperué, W., & Lanzarini, L. C. (2014). Keyword extracting using auto-associative neural networks. *XX congreso argentino de ciencias de la computación* 2014.

Bahdanau, D., .Cho, K., .& Bengio, Y. (.2014). “Neural machine translation by jointly learning to align and translate,” ArXiv Prepr. ArXiv 14090473.

Barzilay, R., & Elhadad, M. (1999). *Using lexical chains for text summarization. Adv. Autom. Text Summ.*111–121.

Beliga, S. (2014). *Keyword extraction: A review of methods and approaches.* Univ. Rij. Dep. Inform. Rij.

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences, 39*(Jul (1)).

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2016). Selectivity-based keyword extraction method. *International Journal on Semantic Web and Information Systems, 12*(3), 1–26.

Berend, G., & Farkas, R. (2010). SZTERGAK: Feature engineering for keyphrase extraction. *Proceedings of the 5th international workshop on semantic evaluation* (pp. 186–189).

Cao, Z., .Li, W., .Li, S., .Wei, F., .& Li, Y. (.2016). "AttSum: joint learning of focusing and summarization with neural attention," *ArXiv160400125 Cs*, Apr.

Caragea, C., Bulgarov, F. A., Godea, A., & Gollapalli, S. D. (2014). Citation-enhanced keyphrase extraction from research papers: A supervised approach. *EMNLP. 14. EMNLP* (pp. 1435–1446).

Chen, K. Y., et al. (2015). Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques. *IEEE/ACM Transactions on Audio Speech and Language Processing, 23*(Aug (8)), 1322–1334.

Cho, T., & Lee, J.-H. (2015). Latent keyphrase extraction using LDA model. *Journal of The Korean Institute of Intelligent Systems, 25*(2), 180–185.

Chopra, S., Auli, M., Rush, A. M., & Harvard, S. (2016). Abstractive sentence summarization with attentive recurrent neural networks. *HLT-NAACL* (pp. 93–98). .

Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K., & Liberman, M. (2002). Corpora for topic detection and tracking. In J. Allan (Ed.). *Topic detection and tracking: Event-based information organization* (pp. 33–66). Boston, MA: Springer US.

Cohan, A., .& Goharian, N. (.2016). "Revisiting summarization evaluation for scientific articles," *ArXiv Prepr. ArXiv160400400*.

Collins, E., .Augenstein, I., .& Riedel, S. (.2017). "A supervised approach to extractive summarisation of scientific papers," *ArXiv Prepr. ArXiv170603946*.

Dalal, M. K., & Zaveri, M. A. (2011). Heuristics based automatic text summarization of unstructured text. *Proceedings of the international conference & workshop on emerging trends in technology* (pp. 690–693).

Das, D., & Martins, A. F. (2007). *A survey on automatic text summarization. Lit. surv. lang. stat. II course CMU, 4*, 192–195.

De la Peña Sarracén, G. L., & Rosso, P. (2018). Automatic text summarization based on betweenness centrality. *Proceedings of the 5th Spanish conference on information retrieval* (pp. 11:1–11:4).

"DMQA." [Online]. Available: https://cs.nyu.edu/%7Ekcho/DMQA/ [Accessed 3 June 2019.].

El-Beltagy, S. R., & Rafea, A. (2009). KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information System, 34*(Mar (1)), 132–144.

El-Kishky, A., Song, Y., Wang, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. *8*, (pp. 305–316).

Ermakova, L. (2012). *Automatic summary evaluation. Rouge modifications. Russ*, 2012.

Fang, C., Mu, D., Deng, Z., & Wu, Z. (2017). Word-sentence Co-ranking for automatic extractive text summarization. *Expert Systems with Applications, 72*(Apr (C)), 189–195.

Fattah, M. A., & Ren, F. (2008). *Automatic text summarization. vol. 37*, World Academy of Science, Engineering and Technology2008.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: A survey. *Artificial Intelligence Review, 47*(Jan (1)), 1–66.

Ganesan, K., Zhai, C., & Han, J. (2010). Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. *Proceedings of the 23rd international conference on computational linguistics* (pp. 340–348). .

García-Hernández, R. A., & Ledeneva, Y. (2009). *Word sequence models for single text summarization.* 44–48.

Gayo-Avello, D., Alvarez-Gutierrez, D., & Gayo-Avello, J. (2004). Naive algorithms for keyphrase extraction and text summarization from a single document inspired by the protein biosynthesis process. *Biologically inspired approaches to advanced information technology, Heidelberger Platz 3, D-14197 Berlin, Germany. vol. 3141. Biologically inspired approaches to advanced information technology, Heidelberger Platz 3, D-14197 Berlin, Germany* (pp. 440–455).

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 19–25).

Gupta, M., & Garg, N. K. (2016). Text summarization of Hindi documents using rule based approach. *2016 international conference on micro-electronics and telecommunication engineering (ICMETE)* (pp. 366–370). .

Haddoud, M., & Abdeddaim, S. (2014). Accurate keyphrase extraction by discriminating overlapping phrases. *Journal of Information Science, 40*(Aug (4)), 488–500.

Haddoud, M., Mokhtari, A., Lecroq, T., & Abdeddaïm, S. (2015). Accurate keyphrase extraction from scientific papers by mining linguistic information. *CLBib@ ISSI* (pp. 12–17). .

Harman, D. (1996). The text retrieval conferences (trecs). *Proceedings of a workshop on held at Vienna, Virginia: May 6-8. 1996. Proceedings of a workshop on held at Vienna, Virginia: May 6-8* (pp. 373–410).

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In: *Proc. of the 52nd annual meeting of the association for computational linguistics (ACL)*.

Hermann, K. M., et al. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems* (pp. 1693–1701). .

Hong, S. G., Shin, S., & Yi, M. Y. (2014). Contextual keyword extraction by building sentences with crowdsourcing. *Multimedia Tools and Applications, 68*(Jan (2)), 401–412.

Huang, C., Tian, Y., Zhou, Z., Ling, C. X., & Huang, T. (2006). Keyphrase extraction using semantic networks structure analysis. *Sixth international conference on data mining (ICDM'06)* (pp. 275–284). .

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 216–223). .

Janin, A. (2003). The ICSI meeting corpus. *Acoustics, speech, and signal processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE international conference on. 1* I–I.

Jo, T. (2003). Neural based approach to keyword extraction from documents. *Computational science and its applications — ICCSA 2003* (pp. 456–461). .

Jo, T., & Lee, J.-H. (2015). Latent keyphrase extraction using deep belief networks. *International Journal of Fuzzy Logic and Intelligent Systems, 15*(3), 153–158.

Kagan, A. (2019). *Lempel-Ziv compressions. Entropy compression methods.* [Online]. Available http://compressions.sourceforge.net/LempelZiv.html, Accessed date: 3 June 2019.

Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D. (2014). Extractive summarization using continuous vector space models. *Proceedings of the 2nd workshop on continuous vector space models and their compositionality (CVSC)@ EACL* (pp. 31–39). .

Kaikhah, K. (2004). Automatic text summarization with neural networks. *Intelligent systems, 2004. Proceedings. 2004 2nd international IEEE conference. 1. Intelligent systems, 2004. Proceedings. 2004 2nd international IEEE conference* (pp. 40–44).

Khan, A., & Salim, N. (2014). A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology, 59*(1), 64–72.

Khatri, C., .Singh, G., .& Parikh, N. (.2018). "Abstractive and extractive text summarization using document context vector and recurrent neural networks," *ArXiv180708000 Cs*, Jul.

Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 21–26). .

Krapivin, M., & Marchese, M. (2009). *Large dataset for keyphrase extraction*.

Kumar, N., & Srinathan, K. (2008). Automatic keyphrase extraction from scientific documents using N-gram filtration technique. *Proceedings of the eighth ACM symposium on document engineering* (pp. 199–208).

Lahiri, S., Mihalcea, R., & Lai, P.-H. (2017). Keyword extraction from emails. *Natural Language Engineering, 23*(2), 295–317.

Li, S.-Q., Du, S.-M., & Xing, X.-Z. (2017). A keyword extraction method for Chinese scientific abstracts. *Proceedings of the 2017 international conference on wireless communications networking and applications* (pp. 133–137).

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *in Text summarization branches out: Proceedings of the ACL-04 workshop. vol. 8*.

Lin, H., & Bilmes, J. (2011). A class of submodular functions for document summarization. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologiesvol. 1. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 510–520).

Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. *Proceedings of the workshop on multi-source multilingual information extraction and summarization* (pp. 17–24). .

Litvak, M., Last, M., & Kandel, A. (2013). DegExt: A language-independent keyphrase extractor. *Journal of Ambient Intelligence and Humanized Computing, 4*(Jun (3)), 377–387 SI.

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. *Proceedings of the 2010 conference on empirical methods in* *natural language processing* (pp. 366–376).

Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *International conference on theory and practice of digital libraries* (pp. 473–474). .

Lopez, P., & Romary, L. (2010b). HUMB: Automatic key term extraction from scientific articles in GROBID. *Proceedings of the 5th international workshop on semantic evaluation* (pp. 248–251).

Lopez, P., & Romary, L. (2010a). GRISP: A massive multilingual terminological database for scientific and technical domains. *LREC 2010*.

Loza, V., Lahiri, S., Mihalcea, R., & Lai, P.-H. (2014). Building a dataset for summarization and keyword extraction from emails. *LREC* (pp. 2441–2446). .

Lynn, H. M., Choi, C., & Kim, P. (2018). An improved method of automatic text summarization for web contents using lexical chain with Semantic-related terms. *Soft Computing, 22*(June (12)), 4013–4023.

Lynn, H. M., Lee, E., Choi, C., & Kim, P. (2017). SwiftRank: An unsupervised statistical approach of keyword and salient sentence extraction for individual documents. *Procedia Computer Science, 113*(Jan (Suppl C)), 472–477.

Ma, S., Sun, X., Lin, J., & Ren, X. (2018). A hierarchical End-to-end model for jointly improving text summarization and sentiment classification. *Proceedings of the 27th international joint conference on artificial intelligence* (pp. 4251–4257).

Mackey, A., & Cuevas, I. (2018). Automatic text summarization within big data frameworks. *Journal of Computing Sciences in Colleges, 33*(May (5)), 26–32.

Malliaros, F. D., & Skianis, K. (2015). Graph-based term weighting for text categorization. *Advances in social networks analysis and mining (ASONAM), 2015 IEEE/ACM international conference on* (pp. 1473–1479). .

Manning, C. D., et al. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations*.

Marr, B. (2019). *Big Data: 20 mind-boggling facts everyone must read.* Forbes. [Online]. Available https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/, Accessed date: 3 June 2019.

Marujo, L. (2015). Automatic keyword extraction on Twitter. *ACL (2)* (pp. 637–643). .

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (pp. 296–297).

Menaka, S., & Radha, N. (2013, 2321-2325). An overview of techniques used for extracting keywords from documents. *International Journal of Computer Trends & Technology, 4*(7), 2321–2325.

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., & Chi, Y. (2017). Deep keyphrase generation. *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 582–592).

Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2016). Automatic keyphrase extraction: an overview of the state of the art. *2016 4th IEEE international colloquium on information science and technology (CiSt)* (pp. 306–313). .

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on interactive poster and demonstration sessions*.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. *EMNLP. vol. 4. EMNLP* (pp. 404–411).

Mittal, V., Kantrowitz, M., Goldstein, J., & Carbonell, J. G. (1999). *Selecting text spans for document summaries: Heuristics hand metrics.* Institute for Software Research.

Moratanch, N., & Chitrakala, S. (2017). A survey on extractive text summarization. *2017 international conference on computer, communication and signal processing (ICCCSP)* (pp. 1–6). .

Nallapati, R., .Zhai, F., .& Zhou, B. (2016). "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents,". Nov.

Nallapati, R., .Zhou, B., .Gulcehre, C., Xiang, B. (2016), "Abstractive text summarization using sequence-to-sequence RNNs and beyond," *ArXiv Prepr. ArXiv160206023*.

Newman, D., Koilada, N., Lau, J. H., & Baldwin, T. (2012). Bayesian text segmentation for index term identification and keyphrase extraction. *COLING* (pp. 2077–2092). .

Ng, J.-P., & Abrecht, V. (.2015). "Better summarization evaluation with word embeddings for rouge," *ArXiv Prepr. ArXiv150806034*.

Nguyen, T. D., & Luong, M.-T. (2010). WINGNUS: Keyphrase extraction utilizing document logical structure. *Proceedings of the 5th international workshop on semantic evaluation* (pp. 166–169). .

Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using latent semantic analysis. *Journal of Information Science, 37*(4), 405–417.

Park, J., Kim, J., & Lee, J.-H. (2014). Keyword extraction for blogs based on content richness. *Journal of Information Science, 40*(1), 38–49.

Patil, M. P. D., & Kulkarni, N. (2014, 42-45). Text summarization using fuzzy logic. *Paragraph, 1*(3), 42–45.

Paulus, R., .Xiong, C., .& Socher, R. (.2017). "A deep reinforced model for abstractive summarization," *ArXiv170504304 Cs*, May.

Pay, T. (2016). Totally automated keyword extraction. *2016 IEEE international conference on big data (Big Data)* (pp. 3859–3863). .

Pay, T., & Lucci, S. (2017). Automatic keyword extraction: An ensemble method. *2017 IEEE international conference on big data (big data)* (pp. 4816–4818). .

Ray Chowdhury, J., Caragea, C., & Caragea, D. (2019). Keyphrase extraction from disaster-related tweets. *The world wide web conference* (pp. 1555–1566).

Reeve, L., Han, H., & Brooks, A. D. (2006). BioChain: Lexical chaining methods for biomedical text summarization. *Proceedings of the 2006 ACM symposium on applied computing* (pp. 180–184).

Rekabdar, B., Mousas, C., & Gupta, B. (2019). Generative adversarial network with policy gradient for text summarization. *2019 IEEE 13th international conference on semantic computing (ICSC)* (pp. 204–207). .

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). *Automatic keyword extraction from individual documents. Text mining.* John Wiley & Sons, Ltd1–20.

Rush, A.M., Chopra, S., .& Weston, J. (.2015). "A neural attention model for abstractive sentence summarization," *ArXiv Prepr. ArXiv150900685*.

Sandhauts, E. (2008). *The New York times annotated corpus.* Philadelphia: Linguistic Data Consortium.

Sankaran, B., .Mi, H., .Al-Onaizan, Y., .& Ittycheriah, A. (.2016). "Temporal attention model for neural machine translation," Aug.

Sarkar, K. (2012). An approach to summarizing Bengali news documents. *Proceedings of the international conference on advances in computing* (pp. 857–862). Communications and Informatics.

Sarkar, K., Nasipuri, M., & Ghose, S. (2012). Machine learning based keyphrase extraction: Comparing decision trees, naïve Bayes, and artificial neural networks. *Journal of Information Processing Systems, 8*(4), 693–712.

Schluter, N. (2017). The limits of automatic summarisation according to rouge. *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, Short Papers* (pp. 41–45). .

Sevilla, A. F. G., Fernández-Isabel, A., & Díaz, A. (2016). *Enriched semantic graphs for extractive text summarizationLNAIvol. 9868*.

Sharan, L. (.2019). "Keyword finder: Automatic keyword extraction from text." [Online]. Available: https://people.csail.mit.edu/lavanya/keywordfinder. [Accessed 3 June 2019].

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computers and Applications, 109*(2), 18–23.

Silber, H. G., & McCoy, K. F. (2000). Efficient text summarization using lexical chains. *Proceedings of the 5th international conference on intelligent user interfaces* (pp. 252–255).

Song, H.-J., Go, J., Park, S.-B., Park, S.-Y., & Kim, K. Y. (2017). A just-in-time keyword extraction from meeting transcripts using temporal and participant information. *Journal of Intelligent Information Systems, 48*(1), 117–140.

Song, S., Huang, H., & Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications, 78*(1), 857–875.

Steinberger, J., & Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. In: *Proc. ISIM'04* (pp. 93–100). .

Sterckx, L., Demeester, T., Deleu, J., & Develder, C. (2018). Creation and evaluation of large keyphrase extraction collections with multiple opinions. *Language Resources and Evaluation, 52*(Jun (2)), 503–532.

Sze, V., .Chen, Y.-H., Yang, T.-J., & Emer, J. (.2017). "Efficient processing of deep neural networks: a tutorial and survey," *ArXiv Prepr. ArXiv170309039*.

Thomas, J. R., Bharti, S. K., & Babu, K. S. (2016). Automatic keyword extraction for text summarization in e-newspapers. *Proceedings of the international conference on informatics and analytics* (pp. 86:1–86:8).

Tixier, A., Malliaros, F., & Vazirgiannis, M. (2016). A graph degeneracy-based approach to keyword extraction. *Proceedings of the 2016 conference on empirical methods* *in natural language processing* (pp. 1860–1870). .

Treeratpituk, P., Teregowda, P., Huang, J., & Giles, C. L. (2010). Seerlab: A system for extracting key phrases from scholarly documents. *Proceedings of the 5th international workshop on semantic evaluation* (pp. 182–185). .

Tsai, C.-I., Hung, H.-T., Chen, K.-Y., & Chen, B. (2016). Extractive speech summarization leveraging convolutional neural network techniques. *Spoken language technology workshop (SLT)* (pp. 158–164). 2016 IEEE.

Tseng, Y.-H. (1998). Multilingual keyword extraction for term suggestion. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 377–378). .

Tsoumou, E. S. L., Lai, L., Yang, S., & Varus, M. L. (2016). An extractive multi-document summarization technique based on fuzzy logic approach. *2016 international conference on network and information systems for computers (ICNISC)* (pp. 346–351). .

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval, 2*(May (4)), 303–336.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *European conference on machine learning* (pp. 491–502). .

Ulrich, J., Murray, G., & Carenini, G. (2008). A publicly available annotated corpus for supervised email summarization. In: *Proc. of AAAI email-2008 workshop*.

Wan, X., & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. *Proceedings of the 23rd national conference on artificial intelligence - Volume 2* (pp. 855–860).

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese broadcast news corpus. *International Journal of Computer Linguistics Chinese Language Processing, 10*(2), 219–236.

Wang, J., Peng, H., & Hu, J. (2006). *Automatic keyphrases extraction from document using neural network. Advances in machine learning and cybernetics*. Berlin, Heidelberg: Springer633–641.

Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., & Du, Q. (2018). A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *Proceedings of the 27th international joint conference on artificial intelligence* (pp. 4453–4460).

Wang, W. M., Li, Z., Wang, J. W., & Zheng, Z. H. (2017). How far we can go with extractive text summarization? Heuristic methods to obtain near upper bounds. *Expert Systems with Applications, 90*(Dec (C)), 439–463.

Waterford Technologies. (22 Feb 2017). *Big Data - Interesting Statistics, Facts & Figures.* Waterford Technologies.

Weerasooriya, T., .Perera, N., .& Liyanage, S.R. (2017). "KeyXtract Twitter model – An essential keywords extraction model for Twitter designed using NLP tools," *ArXiv170802912 Cs*, Aug.

Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation, 1*(Jun (2)), 270–280.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: practical automatic keyphrase extraction. *Proceedings of the fourth ACM conference on digital libraries* (pp. 254–255). .

Wong, K.-F., Wu, M., & Li, W. (2008). Extractive summarization using supervised and semi-supervised learning. *Proceedings of the 22Nd international conference on computational linguistics – Volume 1* (pp. 985–992).

Wozniak, R. (1999). *Classics in psychology 1855-1914: Historical essays.* Thoemmes press.

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

Wu, K., Shi, P., & Pan, D. (2015). An approach to automatic summarization for Chinese text based on the combination of spectral clustering and LexRank. *Fuzzy systems and knowledge discovery (FSKD), 2015 12th international conference on* (pp. 1350–1354). .

Wu, X., Du, Z., & Guo, Y. (2018). A visual Attention-based keyword extraction for document classification. *Multimedia Tools and Applications, 77*(Oct (19)), 25355–25367.

Wu, Y. B., Li, Q., Bot, R. S., & Chen, X. (2005). Domain-specific keyphrase extraction. *Proceedings of the 14th ACM international conference on information and knowledge management* (pp. 283–284).

Xie, F., Wu, X., & Zhu, X. (2017). Efficient sequential pattern mining with wildcards for keyphrase extraction. *Knowledge-Based Systems, 115*(Jan), 27–39.

Yang, M., Liang, Y., Zhao, W., Xu, W., Zhu, J., & Qu, Q. (2018). Task-oriented keyphrase extraction from social media. *Multimedia Tools and Applications, 77*(Feb (3)), 3171–3187.

Yang, S., Lu, W., Yang, D., Li, X., Wu, C., & Wei, B. (2017). KeyphraseDS: Automatic generation of survey by exploiting keyphrase information. *Neurocomputing, 224*(Feb.), 58–70.

Yao, K., Peng, B., Zweig, G., Yu, D., Li, X., & Gao, F. (2014). Recurrent conditional random field for language understanding. *Acoustics, speech and signal processing (ICASSP), 2014 IEEE international conference on* (pp. 4077–4081). .

Ying, Y., Qingping, T., Qinzheng, X., Ping, Z., & Panpan, L. (2017). A Graph-based approach of automatic keyphrase extraction. *Procedia Computer Science, 107*(Jan), 248–255.

You, W., Fontaine, D., & Barthès, J.-P. (2013). An automatic keyphrase extraction system for scientific documents. *Knowledge and Information Systems, 34*(Mar (3)), 691–724.

Yousefi-Azar, M., & Hamey, L. (2017). Text summarization using unsupervised deep learning. *Expert Systems with Applications, 68*(Feb (C)), 93–105.

Yu, L., & Ren, F. (2009). A study on cross-language text summarization using supervised methods. *2009 international conference on natural language processing and knowledge engineering* (pp. 1–7). .

Zhang, Q., Wang, Y., Gong, Y., & Huang, X. (2016). Keyphrase extraction using deep recurrent neural networks on Twitter. *EMNLP* (pp. 836–845). .

Zhang, Y., Chang, Y., Liu, X., Gollapalli, S. D., Li, X., & Xiao, C. (2017). MIKE: Keyphrase extraction by integrating multidimensional information. *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 1349–1358).