Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 75-82 recibido 28-04-2020 revisado 04-06-2020 aceptado 04-06-2020

Relevant Content Selection through Positional Language Models: An Exploratory Analysis

Selección de Contenido Relevante mediante Modelos de Lenguaje Posicionales: Un Análisis Experimental

Marta Vicente, Elena Lloret
Department of Software and Computing Systems
University of Alicante, Spain
{mvicente,elloret}@dlsi.ua.es

Abstract: Extractive Summarisation, like other areas in Natural Language Processing, has succumbed to the general trend marked by the success of neural approaches. However, the required resources—computational, temporal, data—are not always available. We present an experimental study of a method based on statistical techniques that, exploiting the semantic information from the source and its structure, provides competitive results against the state of the art. We propose a Discourse-Informed approach for Cost-effective Extractive Summarisation (DICES). DICES is an unsupervised, lightweight and adaptable framework that requires neither training data nor high-performance computing resources to achieve promising results.
Keywords: Summarisation, Positional Language Models, Discourse Semantics

Resumen: Como muchas áreas en el ámbito del Procesamiento de Lenguaje Natural, la generación extractiva de resúmenes ha sucumbido a la tendencia general marcada por el éxito de los enfoques de aprendizaje profundo y redes neuronales. Sin embargo, los recursos que tales aproximaciones requieren – computacionales, temporales, datos – no siempre están disponibles. En este trabajo exploramos un método alternativo basado en técnicas estadísticas que, explotando la información semántica del documento original así como su estructura, proporciona resultados competitivos. Presentamos DICES, un método no supervisado, económico y adaptable que no necesita recursos potentes ni grandes cantidades de datos para lograr resultados prometedores respecto al estado de la cuestión.
Palabras clave: Resúmenes automáticos, Modelos de Lenguaje Posicionales, Semántica del Discurso

1 Introduction

The explosive growth of data that today's society witness, puts in the front line the research and development of suitable technologies that facilitate not only the access to such data deluge, but its comprehension. To this effect, summarisation techniques become a crucial resource, aiming at facilitating information access and understanding by condensing data with no loss of meaning (Nenkova and McKeown, 2011).

Deep Learning (DL) approaches have become increasingly popular in most of the Natural Language Processing (NLP) tasks, also in text summarisation, with competitive results and promising developments[1] for enabling transformation in industry and academia. However, today shortcomings of DL technologies raise concerns at different levels depicting a landscape of affected scenarios: small companies that cannot access huge amount of data, direct absence of such volume of data due to the specificity of the problem (e.g. some medical realms, organisational documentation, etc.) even the complex processing involved in DL, which may be costly not only in terms of computational resources, but in environmental impact (Strubell, Ganesh, and McCallum, 2019). Drawbacks of this technology, as it exists today, which highlight the need to explore alternative methodologies and motivate our current investigation.

Refocusing attention on statistical methods, in this paper we present a Discourse-

[1] `nlpprogress.com` is a repository to track the progress in NLP, for the most common tasks.

ISSN 1135-5948. DOI 10.26342/2020-65-9

Informed approach for Cost-effective Extractive summarisation (DICES), and examine its performance in the task of single document extractive summarisation (SDS). Since it is conceived as an unsupervised method that does not require human annotation or intervention neither copious amounts of data to provide good results, our approach represents a more digitally inclusive tool for public and private sector organisations that have tighter budgets for accessing Information and Communication Technologies. In this sense, organisations with small scale production of digital data could benefit from this lightweight mechanism to obtain the desired summaries from their different types of information. Our approach is feasible with less resources than would be necessary for deep or machine learning perspectives.

Central to our approach is the fundamental integration of a specific type of language model known as Positional Language Model (PLM), which has proven to be useful and cost-effective in different areas of NLP, such as information retrieval (Boudin, Nie, and Dawes, 2010) or language generation (Vicente, Barros, and Lloret, 2018). Taking into account the positional information of relevant elements within the document, we outline the significance of understanding and processing the text not as a simple set of words but as a succession of messages, whose full meaning must be accessed beyond the sentence level, at the discourse dimension. We shape our methodology to incorporate semantic information gathered from the document, instead of merely using words. Thereby, rather than simply relate to word counts, we aim at overcoming the limitations of the *bag of words* approaches by considering the original texts as structurally meaningful sources of information in a discourse-informed process, showing that such strategy have a positive impact on both the selection of content and the consequent generation of quality summaries.

In short, our contributions are: 1) we define a discourse-informed statistical model which incorporates semantics from the original text, revealing an improvement of the resulting summary; 2) we implement an unsupervised, lightweight and adaptable framework, simple yet effective and 3) we conduct a series of experiments over standard benchmarks and, with no heavy load of computational resources, empirically verify the effectiveness of our approach.

2 Related Work

As our approach is primarily designed for extractive summarisation, perspective that includes salient sentences from the original text without modification, we focus specifically on representative methods within this context, for brevity. We also include related work that underline the importance of discourse, as this is a core aspect for DICES.

Neural approaches have become mainstream research in recent years, where summarisation is tackled as a classification problem which establishes sentence appropriateness. Reinforcement learning (Wu and Hu, 2018; Chen and Bansal, 2018) or encoder-decoder architectures (Cheng and Lapata, 2016; Nallapati et al., 2016) are common, and research into new combinations grows constantly. As opposed to DICES, these approaches still need huge amounts of training data, which is not always available.

Whereas a large part of existing research relies on occurrence frequency, a few studies have focused on including discourse and semantics in their approaches. (Liu, Titov, and Lapata, 2019) proposed a structured attention architecture to induce trees while, in the case of (Liu and Chen, 2019) and (Hirao et al., 2013), they relied in Rhetorical Structural Theory (Mann and Thompson, 1987). The difference with our approach is that these systems include an expensive linguistic component that requires dependency parsing or rhetorical analysis to obtain the relation between the units of the document. DICES represents the semantics and structure of the components from a statistical perspective which imply shallow features and resources.

3 DICES Approach

In this Section we first explain the statistical foundations of DICES to later describe how a middle representation of the document is built upon the PLMs, serving as basis for the method to obtain the required summary.

The fundamental assumption here is that the better the understanding of the original text, the more informative the summary becomes. And only considering the text as a structured discourse, whose semantic elements coherently relates to each other, can that understanding be leveraged.

3.1 Positional Language Models

The basic idea behind this model is that for every position i within a document D, it is possible to calculate a score for each element w that belongs to the document's vocabulary. This value displays the relevance of w in a precise position, based on the element's distance to other occurrences of the same element throughout the document. The closer the elements appear to the position being evaluated, the higher the score obtained. This behavior allows the model to express the significance of the elements considering the whole text as their context, rather than being limited to the scope of a single sentence. In this manner, one PLM is computed for each and every position of the document, which can be formulated as follows:

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \quad (1)$$

where *c(w,j)* indicates the presence of w in position j; $|D|$ refers to the length of the document D; V, the vocabulary and *f(i,j)* is the propagation function rating the distance between i and j.

3.2 PLM for summarisation

Having explained the model foundations, and thus the PLM module, a more detailed procedure needs to be designed to adapt the model to the task of summarisation. This procedure comprises three stages. First, we need the **definition of the vocabulary** as a parameter for the PLM module. From this stage, we obtain a representation of the text that involves both the vocabulary and the positions of its elements. Second, we **create a seed**, i.e., a set of words that can be relevant for the text and whose constitution depends on the corpus of origin itself. Finally, the processing of the PLM against the seed allows us to establish scores for the text elements, which will be transformed into a **ranking of sentences** from which the highest scored ones could be selected to produce the final summary up to a specific length.

3.3 The Vocabulary Definition

First, it is necessary to define which type of elements will constitute the vocabulary for the PLM. A straightforward approach would be to select the words as they appear in the text. However, this could lead to unnecessary repetitions. Alternatively, lemmas could be selected to obtain more effective results and, at a deeper and more comprehensive level, we could also consider more abstract forms, as identifiers for synonyms or deeper semantic or syntactic constructs.

In our current configuration, the vocabulary is composed of the synsets corresponding to nouns, verbs and adjectives, together with the named entities (NE) that appear along the text. This decision aligns with our semantic goal, that in this case was to capture the meanings, and the semantic information they convey. Freeling (Padró and Stanilovsky, 2012) is an open-source tool which provides several layers of linguistic analysis. We use it also to obtain synsets from WordNet (Kilgarriff and Fellbaum, 2000).

3.4 Seed Creation

A seed is then created, which must contain elements that allow us to dislodge the irrelevant parts of the discourse. The process of creation can begin with a sentence or with a set of words, that need to be analyzed with the same tools as the source text. A second vocabulary is then built from those.

Let V denote the source vocabulary, with elements $\{w_1,..,w_{|V|}\}$, and V_s the vocabulary extracted from the seed, a filter vector F is generated with as many positions as elements V has. If the element w_i from V belongs to V_s, then *F[i] = 1* ; *F[i] = 0* otherwise. Now it is possible to obtain a Score Vector (SC) with values for every position j:

$$SC[j] = \sum_{w \in V} P(w \mid j) \times F \quad (2)$$

The general vocabulary has been reduced to relevant vocabulary and, thanks to the PLMs, for each position of the document we dispose of a reference value. SC would become a detector of important areas, with maximum values when the accumulation of relevant elements given by PLMs is higher.

3.5 Ranking and Selection

From the SC, we are able to obtain the positions of interest within the document. From those, we retrieve as candidate sentences the sentences where these positions belong to, and subsequently obtain a value S_{score} for each of those sentences:

$$S_{\text{score}} = \sum_{i \in S} SC[i] \quad (3)$$

with S being the sentence to be scored, and i indicating the positions within the document for that sentence. Since it is possible that the position belongs to a very short sentence, and considering that the value in each position comes from its context, it would be pertinent to include the neighboring sentences. In this way, we introduce a parameter in the processing that allows us to select if we want to recover only the sentence that includes the position, or also its neighbors.

Typically, a parameter needs to be set in order to determine the length of the summary required. We sequentially select the highest scoring sentences, until the established length, to gather the best set of sentences for the resultant summary.

4 *Datasets, Experiments and Evaluation*

In order to evaluate DICES, several experiments were conducted over different corpus. Although the process is similar for all them, some aspects had to be adapted. Next, the datasets are introduced together with some details on the implementation. We finish with some notes on the evaluation metrics.

4.1 Datasets Description

The datasets selected to evaluate DICES were chosen for their renown, to enable a quality comparison with previous systems. In this section, we describe these datasets.

DUC 2002 Some of the most popular datasets used to address summarisation tasks come from the Document Understanding Conferences[2]. (DUC). Among others, DUC2002 includes a task aimed at SDS, whose goal was to built summaries from one English news article of 100 words at most.

CNN/DailyMail We selected a second corpus, the CNN/DailyMail (CNNDM) (Hermann et al., 2015), which is more recent than DUC2002 and is widely used to evaluate DL approaches from both extractive and abstractive perspectives. Apart from the abstractive gold standard in the form of highlights, some authors have created purely extractive model summaries. Particularly, the authors in (Cheng and Lapata, 2016) tagged with a label 1 the sentences from a document which should appear in a gold standard summary, and made the resulting dataset available[3].

One particularity of the corpus is that the documents are presented in an anonimized mode (we call it M for *mentions*), so that the entities appearing in the text are substituted by an identifier or mention. Then, along with the text, a list of the correspondent entities is provided. We processed the corpus to obtain a non-anonimized version (we call it E for *entities*), with the entities in their place. In this manner, our evaluation is conducted on both versions, M and E.

Although the version processed by the authors contains more than 280K documents, we selected a portion of the test documents to evaluate our system, since the strength of DICES does not rely on the amount of examples. Originally, the test set is comprised of 10,397 Daily Mail and 1,093 CNN documents. We took the 1,093 documents from CNN and randomly selected the same amount from Daily Mail, thus creating a smaller, but balanced dataset of 2,186 documents. For this collection, we estimate an average of 17 sentences per document labeled with 1 (following (Cheng and Lapata, 2016)'s version), and 4 highlights per document, also on average.

4.2 Implementation details

As introduced in Section 3, three stages were required to create the final summary: the vocabulary definition, the seed creation and selection of sentences after ranking them. The constitution of the seed is specific for each corpus. DUC2002 provides one headline per article. In this case, the headline, the first sentence or the combination of both would work as seed for the summaries. Regarding CNNDM, headlines are not provided but entities or mentions are. Therefore, the seed could consist of the first sentence of the document, the entities/mentions provided or their combination. For both corpora, we experimentally determined that the best strategy was to select as seed the combined option, for which results are reported.

4.3 Evaluation

We adopt ROUGE (Lin, 2004) for performing the evaluation, a recall-oriented measure which has become one of the most common metrics in summarisation. Summaries were evaluated taking into account ground truth

[2] duc.nist.gov

[3] github.com/cheng6076/NeuralSum

models and their relation with the system summaries in terms of n-gram overlap.

Regarding the gold-standard summaries, DUC2002 provided up to 4 model summaries for each single document and CNNDM documents are paired with a set of abstractive highlights, that serve as reference summary.

Besides, we created a pure extractive gold summary for CNNDM, taking into account sentences labeled as 1, provided by (Cheng and Lapata, 2016). The gold summary was exclusively used to examine DICES capablility of efficiently retrieving important information from a document. We discuss this issue in the next section.

5 Results and Discussion

To explore and evaluate the effectiveness of our approach, the following analyses were conducted: 1) we used the labeled CNNDM corpus to establish the system's ability to retrieve relevant sentences and, 2) we applied our system to the SDS task.

We next present our results and compare them with several state-of-the-art models. We found out that ROUGE unigram (R1) and bigram (R2) overlapping were usually reported by other systems, but occasionally longest subsequence overlap (RL) was also included. Furthermore, some works used F-score and some employed recall. Taking into account this diversity, and in order to get the clearest idea of our system's performance, we have included in the comparison all the significant approaches, reporting on the measure required in each comparison.

5.1 Relevant Sentence Retrieval

The average number of sentences with label 1 on the selected subset of the CNNDM corpus was 17. In compliance with this restriction, summaries limited to that length were obtained using DICES. These summaries were evaluated against the pure extractive gold summary to prove that our approach successfully detected the sentences where the relevant information in the article resided. R1, R2 and RL are presented in Table 1, both for the anonimized (M) and non-anonimized (E) versions of the corpus.

The results show our model's success in retrieving relevant information. The balance between recall and F-score also indicates that the recovered elements are significant, in the different n-grams modalities. However, the outcomes obtain a much higher value than those achieved when an abstract summary is used as reference. Therefore, in this case, we do not compare them with the other systems.

CNNDM	%	R1	R2	RL
M	R	83.18	74.54	81.03
	F	72.00	63.96	70.01
E	R	80.72	71.01	78.27
	F	71.17	61.93	68.86

Tabla 1: DICES evaluation against the pure extractive gold summaries from CNNDM—anonimized (M) and non-anonimized (E)

5.2 System comparison

In this section we compare our experimental results with state-of-the-art systems[4]. Additionally, some baselines are included to provide evidence of our achievements.

5.2.1 DUC2002 Evaluation

We evaluated DICES against several summarisation approaches and report the results in Table 2. The top part of the table exhibits systems that reported recall. The best performing system for the competition *BestDuc02* and the Lead baseline the organisers provided are included. The Lead baseline, taking as summary the first 100 words, relies on the assumption that in news genre, the relevant information is located firstly. Although this baseline was only surpassed by 3 systems, it is highly genre-dependent and does not consider semantic knowledge, contrary to our approach, which is easily adaptable to other genres and domains.

Additionally, taking into account that graph-based and statistical methods represent a common ground within extractive tasks, we included results from LexRank (Erkan and Radev, 2004), a popular graph-based technique that uses the PageRank algorithm, and implemented two baselines based on frequency counts that constitute popular references among statistical approaches, performing a *bag of words* strategy: *Tfidf* (term and inverse document frequency involved), and and *Tfisf*, which is a variation of the former, considering the inverse sentence frequency, instead *idf*.

Systems reporting F-scores are placed at the bottom of Table 2. To the best of our

[4]Results taken from the respective literature. Only for Pointer-Gen in CNNDM task, the code available in `github.com/abisee/pointer-generator##looking-for-pretrained-model` was run.

Duc2002	R (%)		
System	R1	R2	RL
Tfisf	37.03	13.39	30.11
Tfidf	38.43	14.39	31.40
Lead	41.13	21.07	37.53
BestDuc02	42.77	**21.76**	38.64
LexRank	43.20	17.94	**38.91**
DICES02	**44.72**	20.02	37.22
	F-score (%)		
System	R1	R2	RL
Pointer-Gen	37.22	15.78	33.90
ChenBansal	39.46	17.34	36.72
DICES02	**45.97**	**20.56**	**38.25**

Tabla 2: Recall and F-score results on the single-document task of DUC2002

knowledge there were no neural approaches reporting recall measure. We only found *ChenBansal* (Chen and Bansal, 2018) system that presents the F-score for results. They propose a reinforcement learning approach for abstractive summarisation and test it on the DUC2002 task, and also present the results for the pointer-generator system *Pointer-Gen* (See, Liu, and Manning, 2017). Although their system is abstractive and ours is not, the model summaries against which all three systems are compared are the same.

5.2.2 CNN/DailyMail Evaluation

As previously stated, our main objective when working with CNNDM was to evaluate DICES' ability to retrieve salient information (Section 5.1), given that the objective is eminently abstractive summarization. Nevertheless, DUC2002 results showed DICES remarkable performance regarding traditional yet competitive models, and moreover, its comparison against neural approaches revealed promising outcomes. We therefore decided to also conduct our summarisation experiments on this corpus, a task that has been usually approached from neural frameworks. Specifically, we carried out the experiments not over the whole dataset but over the subset of 2,186 documents that we had previously studied. These results may not be strictly comparable due to this size factor, but they certainly give us an idea of the potential of our approach. Table 3 summarises the results.

It appears in the Table a baseline that was created taking the first four sentences of each document, as this was the average length of the highlights for that subset. We built a baseline for each version of the subcorpus: the anonimized (M) and non-anonimized (E) one.

CNNDM - E	Non-anonimized		
System	R1	R2	RL
Tf-Isf	27.97	8.37	22.84
Tf-Idf	30.12	9.68	24.64
BL-4sent	35.56	13.89	31.53
Pointer-Gen	35.64	15.08	32.52
LiuLapata_19	**43.85**	**20.34**	**39.90**
DICES-E	34.46	13.09	28.19
CNNDM - M	Anonimized		
System	R1	R2	RL
Tf-Isf	31.00	10.05	25.59
Tf-Idf	32.77	11.25	27.09
BL-4sent	38.02	15.71	33.67
WuHu_18	**41.22**	**18.87**	**37.75**
DICES-M	36.78	14.62	30.27

Tabla 3: F-score (%) results for CNNDM-E and CNNDM-M, against highlights

We have also included the scores from the state-of-the-art systems: *LiuLapata* (Liu and Lapata, 2019) for E and *WuHu* (Wu and Hu, 2018) for M. Finally, we show the score for the provided model of *Pointer-Gen* on our corpora. Although we do not beat its performance, our score is considerably close.

Compared to state-of-the-art systems, the significant differences between the results may be caused 1) by the variation in the amount of data being processed in each case, 2) by the extractive condition of our approach against the other abstractive systems. We also conducted the test with the *Tf-Idf* and *Tf-Isf* set ups, also performing with them extractive summarization. As in the other tasks, DICES performs better than those approaches consistently.

We conducted one last experiment in order to better understand the performance and possibilities of DICES. As mentioned above, our subset of 2,186 documents was not strictly comparable with the state of the art. Nevertheless, we found an experiment in (Cheng and Lapata, 2016) which evaluates their extractive approach on 500 samples from CNNDM, with the highlights paired to the documents as gold standard. We randomly extracted the same number of articles from our data and performed a similar evaluation. The results (F-score) are reported in Table 4, and indicate a substantial improvement as least of 54 %. Nevertheless, we think it would be interesting to evaluate DICES exactly in the same set of documents.

CNNDM	F-score (%)		
500 docs	R1	R2	RL
ChengLapata	21.20	8.30	12.00
DICES-E	**34.14**	**12.83**	**28.05**
	(+61 %)	(+54 %)	(+133 %)

Tabla 4: F-score results computed on a random CNNDM subsample (improvement in brackets)

6 Discussion

The results obtained both with DUC2002 dataset, and in the evaluation of the system's capacity to recover relevant sentences, demonstrate the effectiveness of DICES in achieving the objectives established, and reinforces our effort on enhancing the semantic structure of the discourse as catalyst for progress in summarisation.

However, the results DICES obtained in some of the evaluation settings were lower than expected, for example, when compared to DL approaches. In this case, the different sizes of data evaluated could well explain the variation in the results or the disparity could be attributed to the fact that some of those systems are trained on different huge datasets and just tested in *smaller* datasets—the ones we use to evaluate our approach, as DUCs. In any case, to better understand the lower performance of DICES in the summarisation of CNNDM we plan to deeper analyse the impact of some factors. An aspect to be considered would be the system's evaluation over the whole CNNDM dataset, to check if in this case size is relevant. It is also worth noting that the corpus highlights are originally abstract summaries. This could imply a disadvantage with respect to our approach whose results may be better contextualised performing a manual evaluation of the resultant summaries. A thorough study on how the distinct constitution of the seed affects the outcomes could also give us more insight on what is causing the discrepancies and a wider scenario to enrich the research.

Moreover, we carried out an analysis on the resulting summaries that allowed us to identify some errors originated in the preprocessing stage. We detected, for example, how punctuation marks, mainly quotes, harmed the language analysers. Besides, the inadequate disambiguation of terms from which the synsets proceed also had an impact on the creation of the vocabulary, either coming from the body or from the seed, and affecting both the size of that vocabulary and its semantic composition, thus having a negative effect on the summaries generated.

Finally, it is worth mentioning recent work on summarisation which outlines the benefits of individually dealing with content selection and realisation (Gehrmann, Deng, and Rush, 2018; Cho et al., 2019). DICES is able to perform these tasks separately due to its modular architecture. The PLM stage presents a basic mechanism to detect salient content within a document by means of a condensed meaning representation. Although in this work we have exclusively tested its performance for extractive summarisation, DICES modules could also be part, for example, of an abstractive pipeline by adding a different realisation module. Additionally, DICES is able to work at multiple granularity levels by focusing on the sentences as a whole, specifically on their semantic constituents or even down to the token level. And this represents a crucial difference regarding common extractive approaches that usually relies in the sentence as their basic unit.

7 Conclusion and Future work

This paper explores a methodology for single document extractive summarisation that exploits positional and semantic information to improve the generation of summaries. A novel model based on statistical grounds is proposed. One of the motivations that led us to devise and test an approach like DICES was to provide a competitive alternative against the general trend that exploits neural networks, in contexts where, for one reason or another, computational and temporal resources or data are less accessible. In general, DICES achieves satisfactory results without the need for a large amount of data, training or computational load, in contrast to more sophisticated DL approaches.

The experiments show the capability of the framework both in detecting relevant areas of the document and in retrieving the appropriate sentences to construct relevant summaries. Its performance was successfully evaluated in creating single document summaries in the news domain for English.

The DICES methodology could easily be adapted to other languages, whenever a linguistic analyser is available. Moreover, due to its unsupervised nature and the flexibility DICES exhibits, it can readily be applied also to

different domains and summarisation modalities as multi-document summarisation, query and user focused summarisation or headline generation.

Acknowledgments

This research results from work partially funded by Generalitat Valenciana (*SIIA* PROMETEU/2018/089) and the Spanish Government—*ModeLang* (RTI2018-094653-B-C22) and *INTEGER* (RTI2018-094649-B-I00). It is also based upon work from COST Action *Multi3Generation* (CA18231).

References

Boudin, F., J. Y. Nie, and M. Dawes. 2010. Positional language models for clinical information retrieval. In *Proc. of EMNLP*, pages 108–115.

Chen, Y.-C. and M. Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proc. of the ACL, Vol. 1*, pages 675–686.

Cheng, J. and M. Lapata. 2016. Neural summarization by extracting sentences and words. In *Proc. of ACL*, pages 484–494.

Cho, S., L. Lebanoff, H. Foroosh, and F. Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proc. of the ACL, Vol. 1*, pages 1027–1038.

Erkan, G. and D. R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Gehrmann, S., Y. Deng, and A. Rush. 2018. Bottom-up abstractive summarization. In *Proc. of EMNLP*, pages 4098–4109.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Hirao, T., Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proc. of EMNLP*, pages 1515–1520.

Kilgarriff, A. and C. Fellbaum. 2000. WordNet: An Electronic Lexical Database. *Language*, 76(3):706.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Liu, Y. and M. Lapata. 2019. Text summarization with pretrained encoders. In *Proc. of EMNLP-IJCNLP*, pages 3721–3731.

Liu, Y., I. Titov, and M. Lapata. 2019. Single document summarization as tree induction. In *Proc. of the NAACL, Vol. 1*, pages 1745–1755.

Liu, Z. and N. Chen. 2019. Exploiting Discourse-Level Segmentation for Extractive Summarization. In *Proc. of the 2nd Workshop on New Frontiers in Summarization*, pages 116–121.

Mann, W. C. and S. A. Thompson. 1987. Rhetorical Structure Theory: Description and Construction of Text Structures. In *Natural Language Generation*. Springer Netherlands, pages 85–95.

Nallapati, R., B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proc. of SIGNLL*, pages 280–290.

Nenkova, A. and K. McKeown. 2011. Automatic Summarization. *Foundations and Trends® in Information Retrieval*, 5(2):103–233.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality FreeLing project developer. *Proc. of LREC*.

See, A., P. J. Liu, and C. D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Proc. of ACL*, 1:1073–1083.

Strubell, E., A. Ganesh, and A. McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proc. of the ACL*, pages 3645–3650.

Vicente, M., C. Barros, and E. Lloret. 2018. Statistical language modelling for automatic story generation. *Journal of Intelligent & Fuzzy Systems*, 34(5):3069–3079.

Wu, Y. and B. Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *AAAI Conf. on Artificial Intelligence*, pages 5602–5609.