Knowledge-guided Aspect-based Summarization

1st Ziqian Luo
Oracle Cloud Infrastructure
Seattle, USA
luoziqian98@gmail.com

***Abstract*—Contextualized pre-trained models, such as BERT [1] and BART [2], have shown great potential in various NLP tasks, pushing the state-of-the-art results to a new level. Although studies have shown that those pre-trained models have captured different kinds of knowledge due to the massive corpus they have been trained on [3], injecting task-specific external knowledge often shows further improvement [4]. Here we choose aspect-based abstractive summarization as a case study and explore two different ways to inject external knowledge into BART. One is through a knowledge graph, the other is through human-defined sequence-level scores. Experiment results show that both methods can get an improvement over vanilla BART.**

***Index Terms*—pre-trained models, knowledge graph**

I. Introduction

Recently, contextualized pre-trained models, such as BERT [1] and BART [2], have brought great impact in the NLP universe. Typically, these models are trained on unlabeled large corpus to get good representations of words, and those representations can be used for different kinds of downstream tasks. Despite the huge success of these pre-trained models, recent works have shown their inability to capture rich knowledge. For example, [5] argue that BERT can only reason about surface form of entity names while failing to capture factual knowledge. [6] found that pre-trained models can only learn limited syntactic information, and explicitly feed linguistic knowledge can further enhance the performance. [7] suggest that pre-trained language models struggle to complete reasoning tasks that require symbolic operations such as comparison and composition. These observations show the necessity to augment pre-trained models with task-specific external knowledge. Here, we take the task of aspect-base summarization as a case study.

There are mainly two ways of augmenting pre-trained models with external knowledge. One is through modifying the standard language modeling objective function to explicitly take into account other kinds of loss [8]. The other way is to directly feed the knowledge representations to the encoder/decoder side [9]. However, since the second method will modify the pre-trained representations, making the whole training process more computationally expensive, we decide to use the first method to conduct our case study.

Aspect-based abstractive summarization systems are typically trained to generate a summary reflecting a target aspect for a given document. It is attracting more and more people due to its significant advantages. First, it can induce latent structure which will make the model more interpretable and the induced structure can be used for other purposes like document segmentation. Second, it can address the information bottleneck of summarization for long documents, which typically have more diverse contents. Third, it can satisfy more flexible demands for users who only care about some specific aspects of a document.

In this work, we investigate two ways to inject external knowledge into BART, one is through a knowledge graph where we reward generated words that are related to the specified topic, the second is through sequence level supervision where we reward the generated summary based on some human-defined sequence scores.

II. Related Work

A. Massively Pretrained Language Models

Since 2018, starting from the publishing of BERT [1], there is a set of language models released in succession. They are pretrained on extremely large text corpora with different architectures, such as GPT-2 [10], BART [2], T5 [11], and so forth. However, there is a heated discussion about whether these language models learn factual knowledge other than superficial linguistic knowledge. [12] created a commonsense dataset and tested the knowledge reasoning ability of BERT without any fine-tuning. They found that BERT is able to get comparable performance with supervised models, and so are viable knowledge bases. However, recent studies suggest that it's hard for models learned in such an unsupervised manner to learn rich knowledge. For example, [5] suggests that those language models cannot capture factual knowledge despite their good performance in reasoning about the surface form of entity names. [13] observed that BERT cannot capture the meaning of negation (e.g., ”not”). [7] found that language models do poorly on tasks that require symbolic operations such as comparison, conjunction, and composition. Therefore, the attempt to inject new knowledge into these existing models is a new trend.

B. Text Summarization

There are two lines of techniques in text summarization, distinguished by the schema to generate text. One is extractive summarization, which selects and concatenates important parts of the text. In contrast, abstractive summarization aims to generate a new summary rather than only relying on copying.

a) Extractive Summarization: There are many traditional ways to work on extractive summarization, e.g., Bayes rules, Conditional Random Fields, Etc. Techniques for extractive summarization are mostly based on a classification approach

979-8-3503-1129-7/23/$31.00 ©2023 IEEE
DOI 10.1109/CCCAI59026.2023.00012

at the sentence level, where the system learns by examples to classify between summary and non-summary sentences. In recent years, extractive text summarization has been conducted using neural models. A typical one of them, [14], uses neural nets to identify the important sentences in the document. Specifically, they label the training data in a machine-learning approach and then extract features as inputs to the neural model. The neural model would learn to rank these sentences.

b) Abstractive Summarization: For abstractive summarization, sequence-to-sequence (seq2seq) models with attention mechanisms have been the base part of abstractive summarization models. [15] is the first one introducing a neural attention seq2seq model with an attention-based encoder and a neural network language model decoder to the abstractive sentence summarization task, which has achieved a significant performance improvement over conventional methods. Another important step is introducing the copy mechanism, [16] proposed a pointer-generator network that implicitly combines the abstraction with the extraction. This pointer-generator architecture can copy words from source texts via a pointer and generate novel words from a vocabulary via a generator. However, with the surge of pre-train language models, some of them, like BART [2] and PEGASUS [17], have achieved impressive performance in abstractive summarization tasks without copy mechanism.

C. Aspect-based Summarization

Aspect-based summarization aims to generate a summary from a source document with only the contents that are relevant to a specific aspect. It first emerged in the consumer feedback domain, where the system extracts information regarding product properties and feedback sentiment from customers [18]. Recently, [19] investigated several methods to induce latent topical information into sequence-to-sequence abstractive summarization frameworks. They have also created a dataset for aspect-based summarization. Our work is closely related to theirs, using the same dataset as theirs. However, different from their work, we attempt to inject external knowledge to guide the model to generate more topical-related words when generating summaries instead of using only the document information.

D. Knowledge Graph

A Knowledge Graph (KG) is a multi-relational graph composed of entities (nodes) and relations (different types of edges). Each edge is represented as a triple of the form (head entity, relation, tail entity), also called a *fact*, indicating that two entities are connected by a specific relation, e.g., (AlfredHitchcock, DirectorOf, Psycho). It has been widely used for representing factual knowledge in downstream applications such as word sense disambiguation [20], question answering [21], and information extraction [22]. In our experiments, we choose ConceptNet [23] as our knowledge graph. It is a multilingual knowledge base representing words and phrases that people use and the common-sense relationships between them. The knowledge in ConceptNet is collected from a variety of resources, including crowdsourced resources (such as Wiktionary and Open Mind Common Sense), games with a purpose (such as Verbosity and nadya.jp), and expert-created resources (such as WordNet and JMDict).

III. Preliminaries

A. Problem Definition

We formalize the problem of aspect-based summarization as below. Given a document X which contains n tokens $x_1, x_2, ..., x_n$ and an aspect a, our system produces a summary Y which contains m tokens $y_1, y_2, ..., y_m$ that satisfy the following requirements: (1) $m < n$, (2) Y only contains contents that are relevant to a.

B. Base Model

We chose BART as our baseline model, aiming to improve its performance by adding different knowledge. BART is a sequence-to-sequence model pre-trained using denoising objectives and has shown superior performance in text-to-text problems such as text summarization and machine translation.

IV. Proposed Method

The overall architecture of our model is shown in Fig. 1. It follows the architecture of BART, and we investigate both knowledge graph and sequence level score as knowledge sources to train the model.

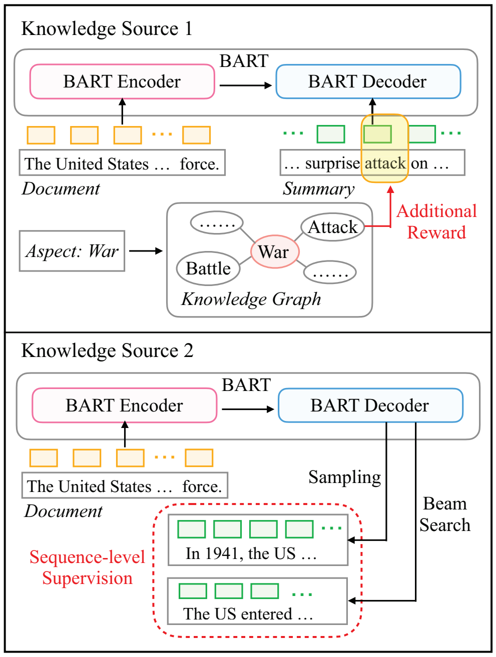

Fig. 1. Proposed methods. We introduce two knowledge sources into BART: one is through a knowledge graph, and the other is through human-defined sequence-level scores.

A. Knowledge into BART

a) Knowledge Source 1: Knowledge Graph: To incorporate the knowledge graph into our model, we make it simple but efficient, adding a word-level supervision loss into the objective function. Specifically, given the input of (aspect, document), we first retrieve all neighbor entities on the knowledge graph with the corresponding entity of the aspect as a bag of words. Then, during training, we give external rewards if the model generates these words. Formally, assume $\mathcal{B}$ is the retrieved bag of words consisting of all neighbor entities with the aspect. We can write the updated objective as

$$\begin{aligned}\mathcal{L} =&(1-\lambda_{KG})\sum_{i=1}^{N} -\log p(y_i|y_0,\ldots,y_{i-1})\\&+\lambda_{KG}\sum_{w\in\mathcal{B}}\sum_{i=1}^{N} -\log p(w|y_0,\ldots,y_{i-1})\end{aligned} \tag{1}$$

where N is the length of the target sequence (the summary) of training examples, λ_{KG} is a hyperparameter representing the weight of this external loss term.

b) Knowledge Source 2: Sentence-Level Score: As we specified above, the most common way to train a summarization model is maximum likelihood estimation. However, exposure bias exists, which is caused by the slightly different procedures between training and inference. In the training phase of maximum likelihood estimation, we give all ground truth tokens to the model and use the model to generate the following token. In contrast, in the inference phase, when we generate the token y_i, we do not have the ground truth of tokens $y_0, y_1, \ldots, y_{i-1}$, so we can only feed the previously generated tokens into the model. In this case, the generation errors would accumulate during the generation process.

Intuitively, when humans evaluate a generated summary, they always take the whole meaning of the sentence instead of taking care of every token. Therefore, any useful sequence-level score can be considered external knowledge for model improvement. Here, we take the schema of self-critical sequence-level policy training, equivalent to applying policy learning as RL, into the sequence training. An external objective can be written as

$$\mathcal{L}_{RL} = (r(y^s)-r(\hat{y}))\sum_{i=0}^{N} -\log p(y_i^s|y_0^s,\ldots,y_{i-1}^s) \tag{2}$$

where y^s is a sampled sequence of the model and $\hat{y}$ is also a generated sequence by taking the token with the maximum probability at every generation step. Similarly, the final loss of training can be written as

$$\mathcal{L} = (1-\lambda_{RL})\sum_{i=0}^{N} -\log p(y_i|y_0,\ldots,y_{i-1}) + \lambda_{RL}\mathcal{L}_{RL} \tag{3}$$

where λ_{RL} is a hyperparameter representing the weight of this external loss term.

V. Experiment

A. Experimental Setting

For all models, we use the Adam [25] optimizer with a linear learning rate scheduling, setting the initial learning rate at 4e-5, and a batch size of 4. The objective is the cross entropy loss with label smoothing factor of 0.1. We fine-tune the models on the whole dataset for 5 epochs. We set a checkpoint at the end of every epoch and finally take the one with the best perplexity on the validation set. We use ConceptNet as our knowledge graph and the averaged ROUGE [26] score as our human-defined sequence-level score. In the generation phase, we use beam search [27] decoding with a beam width of 10 as our decoding strategy for all models.

B. Dataset

[19] synthesized a dataset, MANews, with data samples of the format (aspect, document, summary) = (a, x, y) from the CNN/Daily Mail (CNN/DM) dataset [24], where x is a multi-aspect document, a is an aspect in a pre-defined aspect set (tv showbiz, travel, health, science tech, sports, and news), and y is a summary of x with regard to the aspect a. They assemble synthetic multi-aspect documents, leveraging the article-summary pairs from the CNN/DM corpus, as well as the URL associated with each article, which indicates its topic category. The basic statistics of this dataset are shown in Tab. I. For more details about this dataset, please refer to the original paper [19].

C. Models

We collect results from several previous models for comparisons, including reported results from previous papers and the numbers from our models.

- **Lead-3**: Extract the first three sentences of the article as its summary. The method is aspect-unaware.
- **PG-Net**: Pointer-Generator Network, a typical method in text summarization, has a copy mechanism during generation. In the prediction of a token, it either generates a token or copies a token from the source text. The method is aspect-unaware.
- **Enc-Attn**: Proposed by [19], where the aspect embedding interacts with the word embeddings in the encoder through attention mechanism.

TABLE I
Basic Statistics of the MANews Dataset

Aspect	Num.	Neighbors in ConceptNet
News	47,432	525
Travel	47,567	783
Health	47,549	378
TV Showbiz	47,432	194
Science Technology	47,469	1,255
Sports	47,251	1,276

- **Dec-Attn**: Proposed by [19], where the aspect embedding interacts with the word embeddings in the decoder through attention mechanism.
- **SF**: Proposed by [19], where the aspect embedding is concatenated to word embeddings in the encoder.
- **Enc-Attn-Extract**: Proposed by [19], an extractive summarization model using the same mechanism to incorporate aspect information as **Enc-Attn**.
- **Dec-Attn-Extract**: Proposed by [19], an extractive summarization model using the same mechanism to incorporate aspect information as **Dec-Attn**.
- **BART**: Our base pre-trained model, a performant model in general abstractive text summarization task.
- **BART+CN**: Our model of injecting knowledge from ConceptNet into the base model.
- **BART+SS**: Our model of injecting knowledge from human-defined sequence scores into the base model.
- **BART+CN+SS**: Our model of injecting both ConceptNet knowledge and human-defined sequence scores into the base model.

D. Results

Results from all models are listed in Tab. II. Overall, we make the following observations. (i) Comparing the results of BART and previous non-pre-trained models, we can see the power of pre-trained models, which significantly outperform all previous models. (ii) Comparing the results of BART plus each kind of knowledge, improvements are observed, which means we successfully inject knowledge into the pre-trained model to get improvements. ii) However, when we combine two kinds of knowledge into the model, it does not show further improvement. Therefore, how to make better combination of knowledge from multiple sources to enhance the model should be a direction for our future research.

E. Result Analysis

Here we conduct a more fine-grained analysis. We first break down the results into different aspects, as shown in Tab. III. As indicated by ROUGE scores, we can see that summarizing some aspects is more difficult than others. Typically, contents relevant to "News" and "Sport" are much easier to summarize than others. We make a further investigation to understand the reason behind it.

TABLE II
RESULTS OF BASELINE MODELS AND OUR PROPOSED MODELS.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3	21.50	6.90	14.10
PG-Net	17.57	4.72	15.94
Enc-Attn	27.50	10.27	25.02
Dec-Attn	27.34	10.05	25.09
SF	28.02	10.46	25.36
Enc-Attn-Extract	30.33	10.92	27.32
Dec-Attn-Extract	33.26	13.79	30.26
BART	39.48	18.61	36.61
BART+CN	40.19	18.96	37.30
BART+SS	**40.43**	**19.30**	**37.52**
BART+CN+SS	40.28	19.03	37.42

We first examine the length (i.e., number of words) statistics of gold summaries with regard to different aspects. The results are shown in Tab. IV. From the length statistics, we can see that those summaries for "News" and "Sport" are generally longer than others, indicating that length does not correlate much with the degree of difficulty in summarizing.

After inspecting the length statistics, we are interested in how diverse the gold summaries are with regard to different aspects. Intuitively, if the gold summary is linguistically more diverse, the generated summary should be more difficult to match. The diversity plot is shown in Fig. ref2. Surprisingly, the gold summaries of News and Sport are much more diverse, indicated by the unique n-gram ratio. It seems contradictory to our common sense, but if we consider it more carefully, some factors could lead to this result. The most obvious one is that the source documents for different aspects can be of different degrees of linguistic diversity.

We then examine how extractive gold summaries are by computing ROUGE scores between gold summaries and original documents. The results are shown in Tab. V. It is obvious that News summaries and Sport summaries are more extractive than others.

TABLE III
FINE-GRAINED RESULTS BROKEN DOWN BY ASPECT.

	Model	ROUGE-1	ROUGE-2	ROUGE-L
News	BART	41.16	20.43	38.31
	+ConceptNet	42.16	21.04	39.13
	+SeqScore	**42.88**	**21.73**	**39.85**
Travel	BART	37.29	14.90	34.03
	+ConceptNet	**38.32**	15.42	**34.97**
	+SeqScore	38.23	**15.56**	34.95
Health	BART	38.26	18.85	15.56
	+ConceptNet	39.02	19.42	**36.36**
	+SeqScore	**39.11**	**19.63**	36.25
TV Showbiz	BART	38.02	17.69	35.06
	+ConceptNet	38.60	18.02	35.75
	+SeqScore	**38.80**	**18.33**	**35.94**
SciTech	BART	38.15	18.39	35.67
	+ConceptNet	**38.69**	18.59	**36.22**
	+SeqScore	38.68	**18.76**	36.19
Sport	BART	44.06	21.40	41.10
	+ConceptNet	44.42	21.24	41.44
	+SeqScore	**44.95**	**21.78**	**41.95**

If our models tend to copy from the source documents, this could account for why generated summaries for these two aspects can achieve significantly higher ROUGE scores. We confirmed this hypothesis by examining how extractive our

TABLE IV
LENGTH STATISTICS OF GOLD SUMMARIES W.R.T. DIFFERENT ASPECT

	News	Travel	Health	TV Showbiz	SciTech	Sport
Mean	58	45	57	44	48	49
Std	19	12	66	14	15	13
Min	15	22	8	6	16	20
25%	44	37	36	36	38	38
50%	55	43	44	43	46	46
75%	67	53	56	51	55	56
Max	159	93	505	106	130	133

TABLE V
ROUGE SCORES BETWEEN GOLD SUMMARIES AND ORIGINAL DOCUMENTS

Aspect	ROUGE-1	ROUGE-2	ROUGE-L
News	**8.74**	**5.42**	**8.52**
Travel	6.67	3.46	6.44
Health	7.21	3.93	7.01
TV Showbiz	6.46	3.59	6.26
Science Technology	6.88	4.14	6.73
Sport	7.41	3.91	7.18

models are. The ROUGE scores between generated summaries and original documents are shown in Tab. VI. Compared to gold summaries, our models are generally more extractive. This suggests that although the models were designed to be ”abstractive” - meaning they should generate summaries using their own language and understanding of the source documents - they still tended to rely on the language and phrasing used in the source documents. This could explain why there was a performance gap between different aspects, as some aspects may have had more specific or distinctive language that the models could more easily copy, leading to higher ROUGE scores.

VI. Future Work

a) Abstractiveness Understanding: The analysis for abstractiveness not only explains the performance gap over aspects but also brings about an interesting research topic. Although those models are called ”abstractive summarization models”, they are not that abstractive. To what degree of abstractiveness can a model be? Is abstractiveness controllable? How can we control the abstractiveness effectively? Those series of questions can be left for future work.

b) Knowledge Combination: We observe that simply combining multiple types of knowledge may not necessarily lead to better performance in the model. Therefore, one possible future direction is to investigate how to effectively integrate multiple types of knowledge to improve the model’s performance. This could involve exploring different ways to combine or weigh the different types of knowledge, identifying which types of knowledge are most complementary, or developing new techniques for incorporating multiple sources of knowledge into the model. By improving the model’s ability to utilize multiple types of knowledge, it may be possible to achieve even better performance than using a single type of knowledge alone.

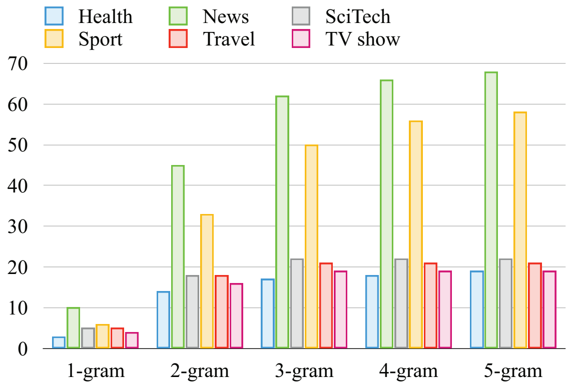

Fig. 2. Diversity of gold summaries w.r.t. different aspect.

c) Broader Scenarios: In this project, as a case study, we incorporated two kinds of knowledge into the pre-trained model on a specific task: aspect-based summarization. In the future, we will find general ways to apply more knowledge to improve the model performance on a diverse set of NLP tasks.

TABLE VI
ROUGE SCORES BETWEEN GENERATED SUMMARIES AND ORIGINAL DOCUMENTS

	Model	ROUGE-1	ROUGE-2	ROUGE-L
News	BART	11.25	10.31	11.22
	+ConceptNet	**11.48**	**10.35**	**11.43**
	+SeqScore	11.36	10.24	11.32
Travel	BART	**10.67**	**9.41**	**10.60**
	+ConceptNet	10.38	8.95	10.38
	+SeqScore	10.47	8.87	10.38
Health	BART	10.37	**9.29**	10.32
	+ConceptNet	**10.54**	9.22	**10.48**
	+SeqScore	10.48	9.15	10.41
TV Showbiz	BART	**10.06**	**8.65**	**9.98**
	+ConceptNet	9.95	8.33	9.86
	+SeqScore	9.87	8.28	9.78
SciTech	BART	10.63	9.74	10.60
	+ConceptNet	10.87	9.82	10.83
	+SeqScore	**11.06**	**10.02**	**11.02**
Sport	BART	**10.96**	**9.08**	**10.84**
	+ConceptNet	10.82	8.71	10.68
	+SeqScore	10.76	8.65	10.62

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “Bert: pre-training of deep bidirectional transformers for language understanding,” in NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186.

[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, et al, “Bart: denoising sequence-to-sequence pre- training for natural language generation, translation, and comprehension,” in ACL 2020, Online, July 5-10, 2020, pp. 7871–7880.

[3] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, et al, “Sparks of artificial general intelligence: early experiments with gpt-4,” unpublished.

[4] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, et al, “A Survey of Knowledge-Enhanced Pre-trained Language Models,” unpublished.

[5] N. Poerner, U. Waltinger, and H. Schutze. “Bert is not a knowledge base (yet): factual knowledge vs. name-based reasoning in unsupervised qa”, unpublished.

[6] A. Lauscher, I. Vulic, E. M. Ponti, A. Korhonen, G. Glavas, “Informing unsupervised pretraining with external linguistic knowledge,” unpublished.

[7] A. Talmor, Y. Elazar, Y. Goldberg, and J. Berant, “olmpics-on what language model pre-training captures,” in TACL 2020, 8:743–758.

[8] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, et al, “Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation,” unpublished.

[9] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, et al, “K-adapter: infusing knowledge into pre-trained models with adapters,” in ACL/IJCNLP 2021, Online Event, August 1-6, 2021, pp. 1405–1418.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, et al, “Language models are unsupervised multitask learners,” unpublished.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, et al, “Exploring the limits of transfer learning with a unified text-to-text transformer,” in J. Mach. Learn. Res., 21:140:1–140:67, 2020.

[12] F. Petroni, T. Rocktaschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, et al, “Language models as knowledge bases?” in EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pp. 2463–2473.

[13] N. Kassner and H. Schütze, “Negated lama: birds cannot fly,” unpublished.

[14] K. Svore, L. Vanderwende, and C. Burges, “Enhancing single-document summarization by combining ranknet and third-party sources,” in EMNLP-CoNLL 2007, pp. 448–457.

[15] A. M. Rush, S. Chopra, and J. Weston. “A neural attention model for abstractive sentence summarization,” in EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pp. 379–389.

[16] A. See, P. J. Liu, and C. D. Manning. “Get to the point: summarization with pointer-generator networks,” in ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp. 1073–1083.

[17] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. “Pegasus: pre-training with extracted gap-sentences for abstractive summarization,” in ICML 2020, 13-18 July 2020, Virtual Event, volume 119, pp. 11328–11339.

[18] J. Zhu, M. Zhu, H. Wang, and B. K. Tsou. “Aspect-based sentence segmentation for sentiment summarization,” in CIKM-TSA 2009, Hong Kong, SAR, China, November 6, 2009, pp. 65–72.

[19] L. Fremann and A. Klementiev. “Inducing document structure for aspect-based summarization,” in ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers, pp. 6263–6273.

[20] R. Nedelchev, D. Chaudhuri, J. Lehmann, A. Fischer, “End-to-end entity linking and disambiguation leveraging word and knowledge graph embeddings,” unpublished.

[21] W. Zhong, D. Tang, N. Duan, M. Zhou, J. Wang, et al, “A heterogeneous graph with factual, temporal and logical knowledge for question answering over dynamic contexts,” unpublished.

[22] B. Sarrafzadeh, A. Roegiest and E. Lank, “Hierarchical knowledge graphs: a novel information representation for exploratory search tasks,” unpublished.

[23] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: an open multilingual graph of general knowledge,” in AAAI 2017, San Francisco, California, USA, pp. 4444–4451.

[24] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, Will Kay, M. Suleyman, et al, “Teaching machines to read and comprehend” in NeurIPS 2015, December 7-12, 2015, Montreal, Quebec, Canada, pp. 1693–1701, 2015.

[25] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” in ICLR 2015, San Diego, CA, USA, May 7-9, 2015.

[26] C. Lin, “Rouge: A package for automatic evaluation of summaries,” in ACL 2004, Barcelona, Spain, July 2004, pp. 74–81.

[27] M. Freitag and Y. Al-Onaizan, “Beam search strategies for neural machine translation,” in NMT@ACL 2017, Vancouver, Canada, August 4, 2017, pp. 56–60