Introducing a Cognitive Model Inductive Bias into Extractive Summarization of Scientific Articles

Ronald Cardenas Acosta (s1987051)
ronald.cardenas@ed.ac.uk

1 Introduction

Automatic summarization is the task of presenting a human user with a short, computer-generated text that retains the most important information from a single or a collection of textual sources. The produced summary is expected to contain information bits, *content units* that are non-redundant and relevant. Content units are also to be presented in a fluent and sensible manner, i.e. in a coherent and cohesive text.

In this proposal, we focus on the problem of selecting informative content units and presenting them in a coherent summary text. To this end, we resort to cognitive theories of human reading comprehension and how they model local and global coherence of a text. Automatic summarization approaches based on psycholinguistic models (Fang et al., 2016; Fang, 2019; Zhang et al., 2016; Lloret, 2011) aim not only to model content relevance but also why this content is relevant in human memory.

In this proposal, we aim to incorporate psycholinguistic mechanisms of memory interaction and attention into a neural summarizer. We plan to focus on theories proposed by Kintsch and Van Dijk (1978) and Kintsch (1988) and their respective computational implementations (Fang, 2019; Zhang et al., 2016; Lloret, 2011). As test bed, we propose to tackle the task of general summarization of long documents, specifically, scientific articles from PubMed and ArXiv. We hypothesize that the bias obtained from structures of memory can be used to explicitly model content selection and planning, and control over granularity of details in the final summary.

We plan to make the following contributions for the next academic year (July 2020 - June 2021):

- Neural architectures that leverage bias from memory trees of content units, pre-calculated from every analysed document. This approach works as a pipeline and relies on an external cognitive model simulator, such as the one proposed by (Fang, 2019)

- Neural architectures capable of deriving structures of memory in an end-to-end manner. This approach accounts for the dynamic nature of memory structure by exploring alternative content unit configurations and all the possible actions necessary to achieve said configurations. This way, we eliminate the dependency on external simulators of cognitive models.

- Comparison of actions taken by our neural models and those taken by humans. We propose to use the task of interactive personalized summarization as a proxy task to infer human actions. Evaluation would be done at the content unit level, based on whether the final combination of modified (or unmodified) content units, i.e. macro propositions, are represented in the final summary. We hypothesize that humans are more efficient at obtaining a final macro proposition than any statistical model, as measured by the number of total actions taken. In addition, we expect humans to represent almost all –if not all– final *gold* macro propositions in the resulting summary.

- As a by-product of the previous item, we plan to introduce `MyNotes`, an interactive tool for computer-assisted personalized summarization. Mechanisms for content presentation and editing are inspired by cognitive theories of text comprehension and implemented by our proposed models.

2 Background

2.1 Cognitive models of text comprehension

Psycholinguistic theories of text comprehension seek to understand the psychological mechanisms interacting when a human reads a text. Most notably, these theories account for limitations in human memory and attention: how much information we can retain and for how long, and which details we preserve and which ones we discard. Theories like Kintsch and van Dijk's, henceforth *KvD* (Kintsch and Van Dijk, 1978), Construction-Integration (Kintsch, 1988), among others (Gernsbacher, 1991; Van den Broek et al., 1996; Zwaan et al., 1995), model human memory as many compartments that, although clearly separated, interact very closely. For example, information from the reader's background knowledge bank may be retrieved to aid comprehension of information found in the text. By modeling these interactions, a cognitive model captures both local and global coherence in the text.

In this proposal, we aim to leverage cognitive structures of human memory in order to introduce an inductive bias for content selection and coherence. We pay special attention to KvD's model of human memory and elaborate as follows.

2.2 Human Working Memory

Working memory is a type of temporal storage equipped with mechanisms for updating and reinforcing information already in place, as well as mechanisms to recall information in permanent memory banks such as long-term memory. Human working memory relies on attention processes to decide what information to update, forget, or reinforce, based on the task being performed. These attention processes are controlled by a central attention module that manages attentional resources among other processes and links the working memory with long-term memory.

2.3 KvD model of Working Memory

KvD proposed a theory that hypothesizes how human working memory works when performing reading and comprehension tasks. The theory states that information in the working memory is organized in two levels of semantic structure, micro-structure and macro-structure. Each level represents semantic units as *propositions* upon which memory operations can be performed.

2.3.1 Proposition

The definition of proposition follows an early definition in cognitive science proposed by Kintsch (1974) and Turner and Greene (1977). We will call propositions that follow this definition *cognitive* propositions. A *cognitive* proposition is the realization of a concept in human working memory. The source of this concept can either be from the text being read or from the user's own personal knowledge. Under this perspective, linguistically-oriented propositions obtained with semantic formalisms are but one kind of cognitive propositions. This definition also implies that the amount of propositions, as well as the level of generality or detail that they represent depends on the knowledge level of the user in the domain of the text being read.

Hence, Kintsch and Van Dijk (1978) deemed sensible to test subjects on domains they were not experts in. The question whether human common-sense reasoning would, in fact, have an effect in proposition granularity is up to discussion. Fang (2019) further formalizes this assumption by stating that the only kind of cognitive propositions stored at the working memory at a given time are *linguistic* propositions, i.e. those extracted from the text by means of semantic formalisms.

2.3.2 Micro-structure

Semantic micro-structure is modeled as a tree of concept units, *propositions*, connected by local coherence criteria. One such criterion is referential coherence which states that a proposition that mentions a referent for the first time superordinates (i.e. is the parent of) any other proposition involving said referent.

Most notably, KvD models the limit on human memory retention by setting a constraint on the number of nodes the memory tree can have at any point in time. In KvD, this limit is said to be 4, and in Fang et al. (2016) the limit is 5 nodes or propositions. The consensus in the cognitive science literature is that this limit is 7 ± 2.

2.3.3 Macro-structure

This structure is also represented as a tree of propositions but it captures discourse organization at the document level and, most importantly, its shape is controlled by the reader's goals in mind. These goals in mind are formally represented by KvD as a *scheme*, i.e. a mental plan the reader builds in order to gather the information needed. KvD hypothesizes that macro-structure can be modeled only in cases where the reading task has a clear, specific goal. Propositions at this level are named *macro* propositions and are not limited by the memory constraint.

2.3.4 Memory Cycle

The reading process is simulated as an iterative process in which the information in the working memory gets updated. As new proposition are detected, some of them are added to the current memory tree and the rest are forgotten. In order to preserve the memory limit constraint, the tree is pruned from nodes deemed to be off-topic.

3 Related Work

3.1 Cognitively-informed Summarization

Recent years have seen the implementation of cognitive models of text comprehension and their direct application to automatic summarization. Zhang et al. (2016) leverages the Construction-Integration theory's capacity of modeling episodic memory to learn the relevance of events in narrative text. Fang et al. (2016) leverages KvD to rank sentences for relevance according to the propositions found in them and how persistent over time these propositions were in the micro-structure of memory. Building upon this work, Fang (2019) incorporates a language generation module and adopts a *extractive-abstractive* approach, i.e. first extracting relevant phrases and then generating an abstractive summary based on them.

It is worth noting that the concept of macro-structure and macro rules was not fully developed until Van Dijk et al. (1983). In this context, Lloret (2011) proposed a system capable of manipulating macro structure and produce a summary.

In this proposal, we aim to build upon the work of Fang et al. (2016) and Lloret (2011), in the sense that we resort on their computational interpretation of micro and macro structure, respectively. In doing so, we seek to leverage inductive bias concerning local and global coherence in a document.

3.2 Contextualized representations of long documents

With the advent of contextualized representations brought by pre-trained architectures (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Yang et al., 2019), the possibility of encoding sentences conditioned on broader contexts became feasible. However, the applicability of this architectures was limited by the number of wordpieces were able to process at a time due to memory and time constraints. For example, `Bert` is limited to 512 pieces.

Hence, early summarization work using transformer architectures had to crop sentences and documents even further so as to fit the data in memory (Liu and Lapata, 2019; Liu, 2019), although notable exceptions also exist (Zhang et al., 2019). As response, a growing body of work is devoting to find ways to simplify the self-attention mechanism in order to process longer sequences efficiently both in time and in memory (Beltagy et al., 2020; Kitaev et al., 2019; Child et al., 2019; Wang et al., 2020).

In this proposal, we aim to summarize long argumentative text –scientific articles– by leveraging linear self-attention architectures (Beltagy et al., 2020). As longer sequences become computationally tractable, the need for a cognitive prior starts gaining relevance.

3.3 Personalization and Interactivity

Summarization is inevitably bounded to subjectivity. In benchmark datasets with more than one reference summary, the low inter-annotator agreement scores is a reminder of the small overlap between what different users consider as *most important* content. Yet current state-of-the-art summarization systems aim to generate content that would fit any user, undermining their usefulness in real-world scenarios. In this scenario, the paradigm of human–computer collaboration offers a viable path to provide the user more control over the content she deems relevant and obtain a tailored summary.

Current work in interactive personalized summarization focused on content exploration rather than sentence post-editing as done in machine translation (Peris et al., 2017; González-Rubio et al., 2012; Isabelle and Chuch, 1997). Strategies on content exploration and selection vary from having the user select spans of text (Leuski et al., 2003) to clicking over predetermined sentences to obtain related sentences with more details (Christensen et al., 2014), or giving the user the option to collapse lists of detail-rich sentences (Shapira et al., 2017). Note that content items in these strategies are fixed and predicted before the interaction begins, i.e. they cannot be edited by the user. In regards to personalized summarization, previous work modeled user preferences through surveys (Berkovsky et al., 2008) and click history as feedback (Yan et al., 2011).

In this proposal we plan to introduce `MyNotes`, a tool that takes a more user-engaging approach to personalization and interaction by giving the user the option of editing selected content by hand or by recommended operations suggested by a cognitively-informed summarizer. In this regard, our approach is most similar to that of (Goldfarb-Tarrant12 et al., 2019), who proposed to assist a human writer during story generation by providing tentative content planning strategies and diversity control.

4 Plan Ahead

We organize our plan for the following year in three stages, which we now elaborate.

Stage 1. Introducing inductive bias from an external cognitive model simulator

In this stage, we tackle the task of extractive *general* summarization of scientific articles. The general idea is to introduce an structured bias that informs the summarizer model how a human would organize information in her working memory as she reads argumentative text. We propose to use the micro-structure of human working memory, as modeled by KvD (Kintsch and Van Dijk, 1978). This micro-structure consists of a tree of propositions that is constantly updated as the document is read. In order to obtain these structures, we resort to the simulator implemented by (Fang et al., 2016). Preliminary experiments showed that (Fang et al., 2016)'s implementation of KvD performed better on long, argumentative text (scientific articles) in comparison with newswire and narrative texts (books).

We formulate the task as a sequence labeling problem where sentence s_i is labeled with $y_i = 1$ if it will be selected for the final summary and labeled with $y_i = 0$ otherwise. We presume the bias obtained from structures of memory will improve the summarizer model capacity to select relevant and coherent content. Therefore, we propose two ways of incorporating inductive bias from memory trees.

Attending Memory Trees. In this approach, a neural network will attend at pre-calculated memory trees while reading a document. When reading sentence s_i, the model would attend at the encoded representation of the working memory at that stage of the cognitive simulation, i.e. tree m_i.

Modeling the dynamic nature of memory structure. Instead of modeling working memory as a sequence of proposition trees, we propose to model the evolution of trees as a sequence of edit-operation actions over nodes. This action-based processing has been successfully applied for parsing (Dyer et al., 2015), lemmatization (Makarov and Clematide, 2018b; Cardenas, 2020), and morphological inflection (Aharoni and Goldberg, 2017; Makarov and Clematide, 2018a). In this line of research, oracle action sequences are obtained when applying iterative algorithms over a sentence input, such as the shift-reduce algorithm for constituency parsing or the transition-based algorithm for dependency parsing. In our case, oracle actions would be obtained from node attachment and forgetting routines in the KvD simulator.

Stage 2. Modeling memory structure evolution in an end-to-end architecture

In this stage we seek to eliminate the reliance on an external KvD simulator for gold trees or gold action sequences. It is worth to note that a certain action sequences might not be the only optimal sequence, and since the model is only exposed to one possible gold sequence during training, we incur in what is called *exposition bias*. In order to remedy this bias, we propose to employ an exploration-based training strategy that would sample the space of all possible action sequences. Previous work (Makarov and Clematide, 2018a) achieves this through Imitation Learning, a reinforcement learning strategy, for the task of morphological re-inflection.

The advantage of this approach is that training is end-to-end without the need of a potentially noisy cognitive simulator. A potential disadvantage is, however, the difficulty of sampling such a huge action space efficiently.

Stage 3. Comparison of structure evolution derived by models versus evolution derived by humans

In this stage, we aim to compare tree-edit actions derived by our proposed models and those performed by humans. However, tree-edit actions cannot be directly obtained from a human working memory without incurring on the phenomenon called *thinking bias?* –ref here–, i.e. an automated, unconscious mental process (e.g. manipulation of the working memory) is interrupted by the task of monitoring said mental process , hence losing the mental structure built so far.

Hence, we propose to use the task of interactive personalized summarization as a proxy task to infer human actions. The task follows under the category of computer-assisted summarization, similar in spirit to recent work on story generation (Goldfarb-Tarrant12 et al., 2019). The idea is to have human users read an article, one chunk at a time, and select excerpts from it to later use as building blocks for a final summary. The size of the chunk would be determined according to the cognitive theory we base the model on, e.g. KvD. The extracted text spans would be editable, mergeable amongst them, or could be generalized–rephrased so that specific details are lost.

Evaluation. Evaluation would be carried over two aspects. First, we want to determine the divergence of memory structure shape at intermediate points during a document reading. An appealing option for intermediate landmarks is the end of paragraphs.

Second, we want to determine how many of the gold content units (e.g. propositions) are present in a user's final summary, according to a reference summary. We hypothesize that humans are more efficient at obtaining final content units –edited or not– than any statistical model, as measured by the number of total actions taken. In addition, we expect humans to represent almost all –if not all– final *gold* content units in the resulting summary.

MyNotes: a tool for personalized summarization As a by-product of the previous study, we plan to introduce `MyNotes`, an interactive tool for computer-assisted personalized summarization. Content is present having KvD theory's memory constraint in mind and the available set of editing operations is based on micro and macro operations defined by KvD. A user would be presented with suggestion of editing actions such as most coherent merges, best items to generalize, and so on. Figure 1 displays a prototype interface of MyNotes.

5 Training needs

The necessary training for stage 1 and 2 consists of learning to use the NLP and Deep Learning tools: Stanford CoreNLP[1] and PyTorch. The author already had training on these tools prior to starting the programme, his knowledge being reinforced again through coursework offered in the courses taken so far.

In regards to theoretical background, we identify the need to learn about specific topics on reinforcement learning, possibly Imitation Learning. The study of cognitive models for text comprehension is in progress but we plan to complement it with material from course `Sentence Comprehension (PSYL11001)`.

In terms of competences regarding stage 3, material from course `The Human Factor: Working with Users (INFR11141)` –complemented with independent research– provided the author with the necessary background knowledge to conduct user studies and to design cognitively-informed, user-friendly interfaces.

[1] `https://stanfordnlp.github.io/CoreNLP/`

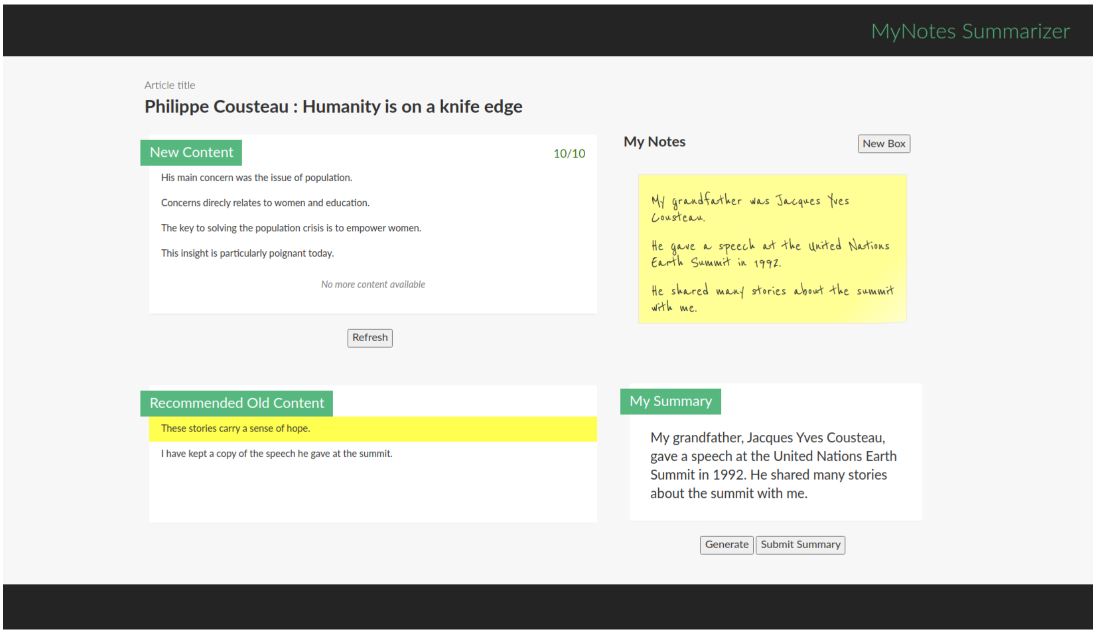

Figure 1: Prototype interface of `MyNotes`, an interactive tool for computer-assisted personalized summarization.

6 Timeline and publication plans

We propose a sensible timeline in sync with the usual submission periods of the most important conferences in the field of NLP and Machine Learning, as shown in Table 1. Currently, stage 1 is under implementation, hence the relatively shorter duration compared to the next stages. We expect to obtain our first results in the following upcoming weeks.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Shlomo Berkovsky, Timothy Baldwin, and Ingrid Zukerman. 2008. Aspect-based personalized text summarization. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 267–270. Springer.

Paul Van den Broek, Kirsten Risden, Charles R Fletcher, and Richard Thurlow. 1996. A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. *Models of understanding text*, pages 165–187.

Ronald Cardenas. 2020. Universal morphological analysis using reinforcement learning. Master's thesis.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Janara Christensen, Stephen Soderland, Gagan Bansal, et al. 2014. Hierarchical summarization: Scaling up multi-document summarization. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 902–912.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Chris Dyer, Miguel Ballesteros, W Ling, A Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Volume 1: Long Papers; 2015 July 26-31; Beijing (China).[place unknown]: ACL; 2015. p. 334-43.* ACL (Association for Computational Linguistics).

Yimai Fang. 2019. *Proposition-based summarization with a coherence-driven incremental model.* Ph.D. thesis, University of Cambridge.

Stage	Estimated time range	Conferences to submit
Stage 1	July-August 2020	AAAI
Stage 2	September-December 2020	ACL, NAACL
Stage 3	January-April 2021	ICLR, NeurIPS

Table 1: Tentative time ranges for each proposed stage of the project, and potential conferences where to submit research output at each stage.

Yimai Fang, Haoyue Zhu, Ewa Muszyńska, Alexander Kuhnle, and Simone Teufel. 2016. A proposition-based abstractive summariser. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 567–578.

Morton Ann Gernsbacher. 1991. Cognitive processes and mechanisms in language comprehension: The structure building framework. In *Psychology of Learning and Motivation*, volume 27, pages 217–263. Elsevier.

Seraphina Goldfarb-Tarrant12, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. *NAACL HLT 2019*, page 89.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254. Association for Computational Linguistics.

Pierre Isabelle and Kenneth Chuch. 1997. Special issue on new tools for human translators. *Machine Translation*, 12.

Walter Kintsch. 1974. The representation of meaning in memory.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, 95(2):163.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Anton Leuski, Chin-Yew Lin, and Eduard Hovy. 2003. ineats: interactive multi-document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 125–128. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Elena Lloret. 2011. *Text summarisation based on human language technologies and its applications*. Universidad de Alicante.

Peter Makarov and Simon Clematide. 2018a. Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877–2882.

Peter Makarov and Simon Clematide. 2018b. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93.

Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2017. Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Ori Shapira, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan, and Ido Dagan. 2017. Interactive abstractive summarization for event news tweets. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 109–114.

Althea Turner and Edith Greene. 1977. *The construction and use of a propositional text base.*

Institute for the Study of Intellectual Behavior, University of Colorado Boulder.

Teun Adrianus Van Dijk, Walter Kintsch, et al. 1983. Strategies of discourse comprehension.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv*, pages arXiv–2006.

Rui Yan, Jian-Yun Nie, and Xiaoming Li. 2011. Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1342–1351. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Renxian Zhang, Wenjie Li, Naishi Liu, and Dehong Gao. 2016. Coherent narrative summarization with a cognitive model. *Computer Speech & Language*, 35:134–160.

Rolf A Zwaan, Mark C Langston, and Arthur C Graesser. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, 6(5):292–297.