Contribution à la lecture automatique à l'aide de réseaux neuronaux profonds

Quentin Grail

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Quentin GRAIL

Thèse dirigée par **Eric GAUSSIER**
et co-encadrée par **Julien PEREZ**, NAVER LABS Europe

préparée au sein du **Laboratoire Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Contribution à la lecture automatique à l'aide de réseaux neuronaux profonds

Contribution to machine reading comprehension with deep neural networks.

Thèse soutenue publiquement le **9 novembre 2021**,
devant le jury composé de :

Monsieur ERIC GAUSSIER
PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur FREDERIC BECHET
PROFESSEUR DES UNIVERSITES, AIX-MARSEILLE UNIVERSITE, Rapporteur
Monsieur ALEXANDRE ALLAUZEN
PROFESSEUR, ESPCI PARIS, Rapporteur
Madame ANNE VILNAT
PROFESSEUR DES UNIVERSITES, UNIVERSITE PARIS-SACLAY, Examinatrice
Monsieur FRANÇOIS PORTET
PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES, Président
Monsieur JULIEN PEREZ
CHERCHEUR, NAVER LABS EUROPE, Invité

CONTRIBUTION TO MACHINE READING COMPREHENSION WITH DEEP NEURAL NETWORKS

by

Quentin Grail

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Laboratoire d'Informatique de Grenoble
NAVER LABS Europe

February 2022

Abstract

Natural Language Understanding is one of the most challenging objectives of Artificial Intelligence. In this dissertation, we describe our contributions related to this field. We investigate several directions that we believe are crucial to build better human language understanding systems. The beginning of the thesis covers essential concepts by proposing a brief history of word representations, machine reading, and automatic text summarization. It describes the diverse objectives that have driven the research community during the last years until the latest revolution of deep learning models for natural language processing.

The first theme developed in this thesis is related to machine reading comprehension or question-answering. Our contributions in this domain are related to 3 aspects: evaluation data, training algorithms, and model design. In this first theme, we propose a question-answering dataset that challenges the relational reasoning competencies of the reader. Then, we discuss our proposed adversarial training algorithm. We designed it to automatically challenge a reading model with corrupted examples in order to improve its performance. Eventually, we describe our work on multi-hop question answering. Machine reading is a vast framework, and novel types of tasks have recently been proposed to evaluate different competencies of reading models. Multi-hop question-answering is one of them and requires the reader to collect multiple pieces of text over a set of documents to answer a question. We believe that this task represents an additional step toward better human understanding models, and we propose our contribution to this domain with an efficient and interpretable deep learning approach.

The burst of deep learning associated with the increasing computational power of modern machines led to impressive achievements in natural language processing. However, these recent architectures tend to be evaluated on tasks that require reading only pieces of text of limited length. The second theme covers in this thesis is related to learning long document representations with state-of-the-art deep learning architectures. We describe our proposition to extend such powerful approaches to tasks that require processing longer documents. We evaluated this proposition on an extractive summarization task of long scientific documents and present some exciting results with minimum adaptation from available works.

Résumé

La compréhension automatique du langage naturel est un défi important de l'intelligence artificielle. Dans cette dissertation, nous décrivons l'ensemble de nos contributions apportées à ce domaine. Nous présentons plusieurs directions que nous pensons cruciales à la construction de meilleurs systèmes de traitement automatique du langage naturel. La première partie de cette dissertation couvre certains concepts essentiels notamment en proposant un historique rapide de la représentation vectorielle de mots ainsi que des tâches de lecture et de résumé automatique de texte. Cette partie décrit certains des principaux objectifs qui ont guidés la recherche durant ces dernières années jusqu'à la récente révolution de l'apprentissage profond appliquée au traitement du langage naturel.

Le premier thème développé dans cette thèse concerne la compréhension automatique de texte au travers de la tâche de question-réponse. Nos contributions dans ce domaine sont liées à trois aspects principaux : les données d'évaluation, les algorithmes d'apprentissage, la construction de nouveaux modèles. Dans ce premier thème, nous proposons un jeu de données de question-réponse permettant d'évaluer les compétences de raisonnement relationnel du système de lecture. Ensuite, nous proposons un protocole d'apprentissage adversarial ayant pour but de générer automatiquement des exemples bruités afin d'améliorer les performances du modèle de lecture. Finalement, nous décrivons nos travaux proposés dans le cadre de question-réponse "multi-hop". La tâche de question-réponse est assez générale et de nouveaux types de questions ont émergés ces dernières années dans le but d'évaluer différentes compétences des modèles de lecture. Les questions "multi-hop" font partie de ces nouvelles directions et nécessite au lecteur de collecter de l'information dans plusieurs

parties de documents afin de répondre correctement à une question. Nous pensons que cette tâche est un pas de plus vers la construction de meilleurs modèles de compréhension du langage et proposons notre contribution au travers d'un modèle de lecture efficace et interprétable.

L'explosion de l'apprentissage profond associé à l'augmentation de la puissance de calcul des machines modernes à conduit à des progrès remarquables dans le domaine du traitement du langage naturel. Cependant, les récentes architectures développées ont tendance à être évaluées sur des tâches nécessitant de lire uniquement des textes de taille relativement modérée. Le deuxième thème couvert dans cette thèse concerne l'apprentissage de représentations de textes longs en utilisant différentes architectures d'apprentissage profond état de l'art. Nous décrivons notre proposition ayant pour but d'améliorer les récentes approches proposées, en les adaptant pour des tâches nécessitant le traitement de documents longs. Nous avons évalué cette proposition sur une tâche de résumé extractif de textes scientifiques et présentons des résultats encourageants ne nécessitant qu'une adaptation minimale des architectures existantes.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context

Language is a fundamental component of human behavior as it serves to communicate and record essential knowledge. Teaching machines to read and understand human language has always been seen as one of the most challenging objectives for Artificial Intelligence. Such an understanding system could lead to many applications such as virtual assistants, automatic text summarizers, translation models, and much more.

However, understanding human language for a machine is not a well-defined concept and still needs to be clarified. Historically, the Natural Language Processing (NLP) community has been focusing on different tasks that were suggested as necessary to understand human language. This includes tasks such as named-entity recognition, syntactic parsing, part-of-speech tagging, and so on. However, even if these tasks are important, they do not guarantee that a model that efficiently solves them will be useful for downstream applications. Moreover, these techniques are, most of the time, individually evaluated and the proposed solution used to be specifically designed for each of them.

A common practice to evaluate how much a human learns from a given piece of text is to ask him questions about it. From this observation, people started to consider such

a question-answering task as a possible framework to evaluate how well a system understands human written documents. Recently, this task of question-answering, or reading comprehension, has rapidly gained in popularity. It is nowadays commonly adopted as a meaningful framework to evaluate how well a system understands a text. First works of this modern area were proposed in 2013 with notably the MCTest dataset (Richardson et al., 2013). The authors strongly believed in the machine reading approach to evaluate the understanding capabilities of a given system but could not find any relevant benchmark to do so. They proposed MCTest, a dataset that will serve as a standard evaluation benchmark to automatically measure human language understanding. Figure 1.1 presents an example from this dataset.

Story: Sara wanted to play on a baseball team. She had never tried to swing a bat and hit a baseball before. Her Dad gave her a bat and together they went to the park to practice. Sara wondered if she could hit a ball. She wasn't sure if she would be any good. She really wanted to play on a team and wear a real uniform. She couldn't wait to get to the park and test out her bat. When Sara and her Dad reached the park, Sara grabbed the bat and stood a few steps away from her Dad. Sara waited as her Dad pitched the ball to her. Her heart was beating fast. She missed the first few pitches. She felt like quitting but kept trying. Soon she was hitting the ball very far. She was very happy and she couldn't wait to sign up for a real team. Her Dad was very proud of her for not giving up.

Question: Who pitched the ball to Sara?
Candidates: A) Her Mom B) Her Sister C) Her Dad D) Her Brother

Figure 1.1: Multi-choice question-answering example from MCTest dataset (Richardson et al., 2013).

Depending on the type of question and type of input document, different competencies of the system are evaluated. The MCTest dataset contains 660 fictional stories associated with multi-choice questions. These questions were designed to be answerable by a 7 years old child.

The question-answering setup has been quite popular and has driven lots of the research community efforts during the last five years, especially with the popular SQuAD dataset (Rajpurkar et al., 2016). It contains factual natural language questions over the most read

Wikipedia articles. It has influenced many research propositions in this field. As a significant milestone, models with *super-human* performance on this task emerged. There is no doubt that these *super-human* performances achieved by reading models are truly impressive and represent one of the biggest achievements of the NLP community of the last years. However, some papers (Sugawara et al., 2018) argue that the task might not be as difficult as we thought and show that a system based on a matching pattern strategy can achieve good performance without any kind of understanding of human language. The first theme develops in this thesis will lie in this question-answering domain and the evaluated competencies. We will first propose a dataset based on restaurant reviews with tasks grounded on human competencies. Then, we investigate adversarial networks and how they can help to automatically complexify the reading task to force the trained model to improve its understanding of language. Additionally, we will investigate the multi-hop question answering task proposed to address some limitations of SQuAD examples. This former task challenges the reasoning competencies of a reader. Indeed, it requires the system to gathered and process information from multiple pieces of text to answer a given question.

With the burst of deep neural networks, Natural Language Understanding went through a huge revolution during the last decade. Especially with the recent Transformer-based (Vaswani et al., 2017) models. However, current standard evaluations such as the GLUE (Wang et al., 2019a) benchmark, which gather performances of systems on a collection of tasks, often required to read only a small piece of text. The second theme of this thesis will focus on bridging the gap between powerful and complex models suitable for small paragraph reading and tasks that require reading long documents. In this part, we will mostly focus on the task of document summarization as it is a challenging evaluation that requires understanding long documents.

1.2 Structure of the Thesis

The work presented in this thesis is related to multiple aspects of Natural Language Understanding and is presented as follow: As a first step, we introduce some background for the thesis. Chapter 2 starts with an overview of word embeddings and the recent evolution of these representations with deep neural networks associated with language modeling objectives. Then we propose an history of machine reading comprehension and details its different possible formulations, namely *cloze style*, *extractive*, *multi-choice* and *free-form* question-answering. We illustrate them with several popular datasets, including ReviewQA (Grail and Perez, 2018), a dataset that we proposed during this thesis. We conclude this section with an overview of the automatic text summarization task to introduce some useful background for Chapter 5

The remaining of the thesis is composed of two parts: In Part I we discuss our contributions related to question-answering. Chapter 3 is based on our work Grail et al. (2018). We propose an adversarial protocol for machine-reading to automatically generate challenging examples and train better models. In the final part of this chapter, we discuss the relationships between this protocol and the recent embedding models based on the masked language modeling task. In Chapter 4 we discuss the multi-hop question-answering task and the challenges that it aims to evaluate. We present our contribution (Grail et al., 2020) on the HOTPOTQA (Yang et al., 2018) dataset.

The second theme developed in this thesis in Part II is related to long-document understanding. This chapter is based on our work Grail et al. (2021). We propose an adaptation of BERT-based models that allows them to read long documents while still benefiting from available pre-trained parameters. We evaluate this proposition on long-document understanding tasks: extractive summarization and long-document matching.

1.3 Articles Written during this Thesis

- Grail, Q., and Perez, J. (2018) ReviewQA: a relational aspect-based opinion reading dataset. In *Conférence sur l'Apprentissage automatique, CAp*.
- Grail, Q., Perez, J., and Silander, T. (2019) Adversarial Networks for Machine Reading. In *ATALA international journal Traitement Automatique des Langues*, volume 59 n°2, pages 77-100.
- Grail, Q., Perez, J., (2019) Training of machine reading and comprehension systems, US Patent US20200134449A1.
- Grail, Q., Perez, J., and Gaussier, E. (2019) Latent Question Reformulation and Information Accumulation for Multi-Hop Machine Reading. *openreview:S1x63TEYvr*
- Grail, Q., Perez, J., and Gaussier, E. (2021) Globalizing BERT-based Transformer Architectures for Long Document Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL

Chapter 2

Background

In this second chapter, we aim to present several concepts that will help the reader with the remaining of the thesis. In the first section, we propose a brief history of word representations. From non-contextual to contextual representations, we describe the evolution of word embedding models during the last decade. In this part, we will also cover the language modeling task and describe the Transformer (Vaswani et al., 2017) architecture. We will see how the combination of these two components became a fundamental building block of modern Natural Language Processing (NLP) architectures.

In the second section of this chapter, we propose an overview of machine reading comprehension (MRC) through the task of question-answering. We begin with a history of the task and present the objectives and challenges that a question-answering system required. Then, we formally describe the four major possible formulations of question-answering datasets and illustrate each of them with multiple examples. In addition, we describe our proposed opinion-reading dataset ReviewQA that has been released at CAp2018 [1] and present the different challenges introduced with it.

We conclude this first section with a presentation of the automatic summarization task to introduce some helpful background for the Chapter 5.

[1] http://cap2018.litislab.fr

2.1 Pre-trained Word Representations

Distributional representations of words (Collobert and Weston, 2008; Bengio et al., 2000; Mikolov et al., 2013b; Devlin et al., 2019), or word embeddings, is nowadays an essential component of any NLP systems. These embeddings embed the semantic meaning of words in dense representations. Continuous representation of words aims to replace traditional symbolic, discrete, or feature engineering representations. A word embedding function maps a word to a high-dimensional vector which could be used either directly to compare the representations or as input for a given system.

These embedding vectors have proven to be usefull for a large collection of downstream tasks such as natural language entailment (MacCartney and Manning, 2008), question-answering (Rajpurkar et al., 2016), machine translation (Zou et al., 2013) and so on. Learning these representations with unsupervised approaches has always been seen as a major objective for the research community. This field has been highly investigated during the last decade, and we would like to report here the latest progress together with the current state-of-the-art approaches.

Successful models for word representation are based on the distributional hypothesis (Harris, 1954). This suggests that words that appear in similar contexts tend to have a similar meaning. This hypothesis is fundamental for many models that have been highly popular in modern deep learning architecture that we describe below.

Pre-trained word embeddings can be divided into two categories: **contextual** and **non-contextual** word embeddings. While the representation of a word is always the same for non-contextual models, contextual embedding systems construct embedding vectors that depend not only on a given word but also on its associated context. To better understand how word embedding models work, we need first to introduce the language modeling task.

Language models are statistical functions that compute the probability distribution over a sequence of input tokens. The probability distribution over a given sequence of T

tokens, $\{w_1, w_2, \ldots, w_T\}$, could be defined with the following equation:

$$P(w_1, w_2, \ldots, w_T) = P(w_1, w_2, \ldots, w_{n-1}) \prod_{t=n}^{T} P(w_t | w_{t-n+1}, \ldots, w_{t-1}) \tag{2.1}$$

Such a model is evaluated with perplexity metrics and requires no labeled data to be trained. Indeed, language models are directly trained from any textual data allowing them to benefits from large collections of text freely available on the internet.

The first neural language model was proposed in 2000 by Bengio et al. (2000). They proposed a scalable neural architecture trained on the Brown corpus and the Associated Press News, two datasets of respectively $1, 1$ and 14 million of words. This architecture was a multi-layer feed-forward neural network followed by a softmax layer. The final layer was in charge of computing the probability distribution of the next word given the past ones. In the last part of this paper, the authors suggested investigating how the feature representations learned by the hidden layers of the language model could be useful for downstream tasks. To the best of our knowledge, this work is one of the first to suggest the idea of using word feature representations of a neural language model in downstream tasks. We will see in the remaining of this first part that this is now a commonly adopted practice.

After this work, Collobert and Weston (2008) and Collobert et al. (2011) demonstrated the effectiveness of neural language models, unsupervisely trained, to improve generalization performance on downstream tasks. In these two papers, the authors showed that the representation learned with the language modeling task improves state-of-the-art approaches on a large variety of tasks such as Semantic Role Labeling, Part-of-speech Tagging, Chunking, or Named Entity Recognition without relying on any *man-made* input features. However, generating word embedding using a language model through a deep neural network was computationally expensive for large vocabulary and not realistic at scale.

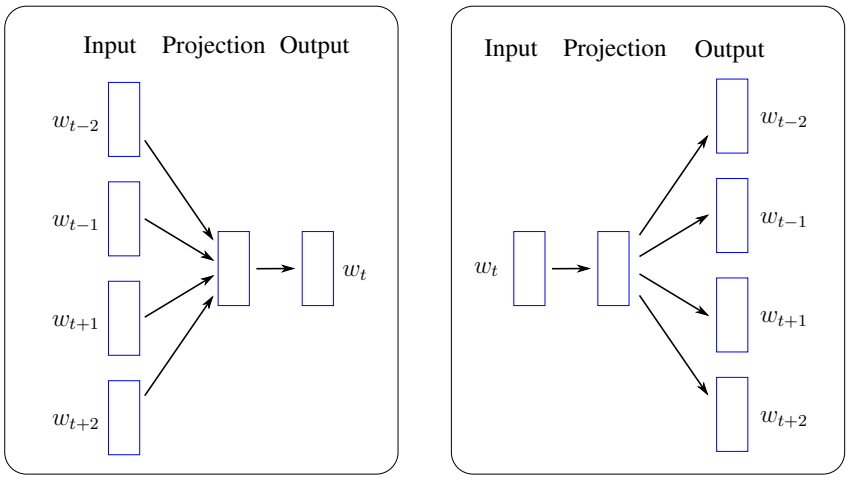

Figure 2.1: Word2Vec Continuous Bag of Word (left) and Skip Gram (right) models.

2.1.1 Non-Contextual Embeddings

Probably the most popular and influential model on word embedding is **word2vec**. It has been proposed proposed in Mikolov et al. (2013b) and Mikolov et al. (2013a). In theses papers, the authors developed two approaches, namely Continuous bag-of-words (CBOW) and Skip-gram, to improve upon Bengio et al. (2000) and Collobert and Weston (2008); Collobert et al. (2011). They demonstrated the possibility of replacing the deep multi-layer feed-forward neural network with a simpler architecture and still computing accurate high-dimensional vectors. It reduced the required computing complexity, thus enabled the algorithm to run faster on much more data. Figure 2.1 presents these two architectures. As presented in the last paragraph, a language model is trained to predict the next word given an history of the past ones. In CBOW, the authors proposed using not only these past words but windows of k words before and k words after the targeted one to make the prediction. In the Skip-gram approach, the algorithm is quite similar but doing the opposite operation. While CBOW aims to predict a target word given its surrounding context, the Skip-gram algorithm predicts the surrounding context of a given source word. These two models have been highly popular and influenced a lot of following works in the domain.

A year later, Pennington et al. (2014) introduced Global Vectors for Word Representation (**GloVe**), a competitive set of pre-trained embeddings. Glove used matrix factorization techniques to build these representations instead of the feed-forward neural networks from wor2vec. Word2vec is considered as a predictive model that learns vectors by optimizing a loss based on the prediction of the target. In contrast, Glove is considered as a count-based model, learning its representations from dimensionality reduction of the co-occurrence counts matrix build from the entire corpora.

A third popular embedding approach that came out three years later is **FastText**. It was proposed by Bojanowski et al. (2017). The major drawback of wor2vec and Glove algorithms is that they could not handle out-of-vocabulary words. If a word has not been seen in the training corpora, then it does not have any embedding representation. FastText tackles this problem by proposing a Skip-gram model based on subword level embedding. Each word is now represented by itself plus its bag of subword embeddings. By doing so, it becomes possible to compute representations for words that did not appear in the training data. This subword embedding strategy is now very popular on state-of-the-art embedding models.

While these approaches have been very popular and improved state-of-the-art results on many downstream tasks, they all have a limiting property. After the training process, each word of the vocabulary is associated to a unique dense vector. This representation is always the same regardless of the context in which this word is used.

2.1.2 Contextual Embeddings

Contextual word embedding models propose to go beyond this unique word representation with an embedding function that depends not only on a given word but also on its associated context. One objective is to compute better representations for polysemous and context-dependent words. In 2015, Dai and Le (2015) proposed to train an autoencoder instead of a language model to improve the general performance of a given model. Inspired by the work in sequence-to-sequence learning from Sutskever et al. (2014b), the

authors suggested a recurrent neural network (RNN) associated with an autoencoding objective. Instead of using the model for translation, they used an autoencoder to reconstruct the original sentence. Similar to language modeling, this task does not require any labeled data, and hidden representations represent words within their context. The key difference with previous approaches is that at the end of the pre-training stage, they do not obtain a fixed vector representation for each word from the vocabulary but a trained model that is able to construct word embeddings from an input sequence. One drawback is that because the word embeddings are not a-priori determined, they cannot be stored in a lookup table. Thus, to compute embeddings of a given sequence, one needs to run the trained model on this sequence.

Later **TagLM** (Peters et al., 2017) emphasis the problem of context-independant word embeddings and proposed to train a LSTM-based (Hochreiter and Schmidhuber, 1997) language model to produce context-sensitive word embeddings. Their method significantly outperformed state-of-the-art approaches on several datasets of chunking and named entity recognition. Two concurrent works, namely **ELMo** (Peters et al., 2018a) and **UlmFit** (Howard and Ruder, 2018) came out shortly after this one. These works are also both based on LSTM layers and language modeling objective. While UlmFit used only unidirectional LSTM layers, ELMo proposed a combination of forward and backward language models allowing the final representation to depend on both left and right context. In addition, ELMo improved by computing the final word representation from multiple layers suggesting that the model learns different properties at multiple levels of the architecture. While ELMo is used to construct features that will be used in a completely different model for downstream tasks, UlmFit came with the idea that the same architecture can be used at pre-training and fine-tuning stages. Only the top layer needs to be adapted for the final task with its different objective, while the remainder of the network can stay identical. We will see in the next paragraph that this strategy is now widely used to design word embedding models.

Much of the progress in this area during the last couple of years came with the Transformer architecture introduced in Vaswani et al. (2017). Let's first remind this architecture

before describing contextual embeddings derived from it.

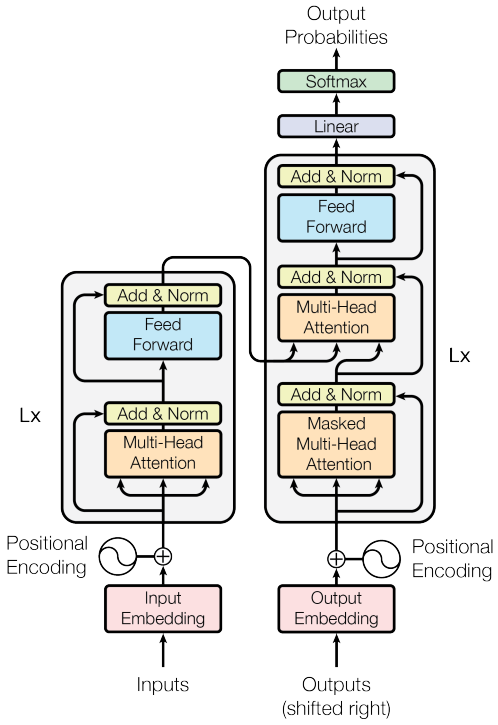

Figure 2.2: The Transformer architecture. Source: Vaswani et al. (2017)

The Transformer is a sequence-to-sequence model originally introduced in the context of machine translation (Vaswani et al., 2017). It consists of an encoder-decoder architecture, each module composed of L identical **transformer blocks**. Figure 2.2 shows the overall architecture. A transformer block is a parametrized function $T_\theta : \mathcal{R}^{d \times n} \rightarrow \mathcal{R}^{d \times n}$ where n is the length of the input sequence and d the dimension of the embedding space.

These blocks highly rely on a multi-head self-attention operation combined with a position-wise feed-forward neural network. The attention mechanism is characterised by the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{2.2}$$

where Q is a matrix representing the query (vector representation of one word in the sequence), K represents the keys (vector representations of all the words in the sequence) and V contains the values (which are again the vector representations of all the words in the sequence).

Instead of computing only a single attention function of dimension d, the authors proposed a multi-head self-attention mechanism that independently computes multiple attention values. We consider a model composed of H independent attention heads. First, for all heads, each key, query, value vector is mapped to a novel projected vector:

$$\begin{aligned} \forall h \in [1, H], \qquad & Q^h = W_h^q Q \\ & K^h = W_h^k K, \qquad \text{where } W_h^q, W_h^k, W_h^v \in \mathcal{R}^{d \times d_k} \\ & V^h = W_h^v V \end{aligned} \tag{2.3}$$

Then we compute the attention describes in Equation 2.2 between these vectors independently on every head:

$$\forall h \in [1, H], \;\; \text{head}_h = \text{Attention}(Q^h, K^h, V^h). \tag{2.4}$$

The final representation is a projection of the concatenation of all heads.

$$O = \text{Concat}(\text{head}_1, \cdots, \text{head}_H) W^o, \;\; \text{where } W^o \in \mathcal{R}^{H d_k \times d} \tag{2.5}$$

As presented in figure 2.2, in the transformer block, the multi-head attention is followed by a 2-layers feed-forward neural network. Each of these operations also has a residual connection and a layer normalization. Compared to classic recurrent neural networks, its biggest benefit comes from its parallelization, which led to an impressive speed up on modern GPU/TPU and the possibility to train deeper models on more data. While RNNs are inherently sequential, the Transformer attention attends to all the tokens of the sequence simultaneously, as described in Equation 2.4. It helps the model to deal with long-term

dependencies issues.

OpenAI's GPT (Radford and Narasimhan, 2018) was the first work to combine the idea of unsupervised pre-training with the Transformer architecture. The decoder of the Transformer architecture is pre-trained on a language modeling signal then fine-tuned on smaller downstream datasets. With minimal adaptation, the pre-trained Transformer improved SOTA of almost 4 points on the GLUE benchmark (Wang et al., 2019a) that gather model performances on a collection of 11 NLP tasks. Compared to previous works, this paper marked a significant difference regarding the computation requirements. While ELMo and other works do not necessitate more than 1 GPU training day, this model was pre-trained during 1 month on 8 GPUs. This is the beginning of a general trend to pre-train bigger and deeper models on massive collections of data.

One of the most influential papers in this area was released shortly after. Devlin et al. (2019) proposed **BERT**, or Bidirectional Encoder Representations from Transformers. They argue that one thing that was missing to GPT is the bidirectionality of the embedding layers. Indeed, GPT is pre-trained on a language modeling signal; thus, the model can only attend to past words to embed the next one. To overcome this limitation, Devlin et al. (2019) proposed the masked language modeling task to pre-trained a Transformer encoder. A certain percentage of input tokens are masked, and the model is trained to retrieve original words using only its context as decision support. In Chapter 3 we will see how this work is related to our proposed adversarial protocol, which also uses a masking strategy to better optimize a given model. In addition to masked language modeling, the authors proposed another self-supervised task which is next sentence prediction. Given a pair of sentences, the model must predict whether they are contiguous or not. The combination of these two pre-training tasks with the Transformer architecture achieved new SOTA with 7 points of improvement over OpenAI's GPT on GLUE.

Starting from this paper a huge variety of Transformer-based language models have emerge using deeper architectures, trained on larger collection of data or using other type of pre-training tasks such as ELMo-Transformer (Peters et al., 2018b), OpenAI GPT2 and GPT3 (Radford et al., 2018; Brown et al., 2020), RoBERTa (Liu et al., 2019b), AlBERT

(Lan et al., 2020), Nvidia Megatron-LM (Shoeybi et al., 2019) and many others. While these modes are extremely powerful, most of them suffer from scalability issues. Because of the self-attention complexity, which is quadratic with the number of input tokens, these Transformer-based language models are pre-trained with sequences of a limited length. For most of them, it is 512 tokens, and it cannot be extended at inference time. To tackle this problem, we propose in Chapter 5 an adaptation of the Transformer layer that can both benefit from available pre-trained languages model and scale to longer sequences without requiring any additional pre-training.

2.2 A Brief Overview of Machine Reading Comprehension

After an introduction to word embedding models, in this second section, we aim to provide to the readers an overview of the question-answering task. A question-answering system is designed to automatically answer natural language questions asked by a human on a given topic. It is a widely used proxy to evaluate the capacities of a model to understand textual data. First models in this domain have been developed in the 1960s/1970s. Notable early works include **Baseball** (Green et al., 1961), a computer program that answers questions phrased in ordinary English about stored baseball data; **LUNAR** (Woods and WA, 1977) which was developed to answer questions about the geological analysis of rocks from the Apollo moon missions; **QUALM** (Lehnert, 1977) defined the theory and proposed a natural language processing system that reads stories and answers questions about what was read. These three early works helped to define the basis of question-answering, but at this time, the proposed approaches were mainly based on handcrafted rules that were written for a specific task. During the following years, there was not a lot of interest in this task, probably due to the lack of computation power and the growing interest of the community for traditional information retrieval systems.

At the beginning of the 2000s, there was a regain of interest for this task, and multiple

scientific events on the topic emerged. The well known Text REtrieval Conference (TREC) started the first question-answering track[2] in 1999, The International Conference on Language Resources and Evaluation (LREC) organized a workshop [3] to discuss a roadmap for question-answering in 2002. Several works have been proposed during this period (Hirschman et al., 1999; Riloff and Thelen, 2000; Voorhees and Tice, 2000; Wang et al., 2000; Cardie et al., 2000). These systems were mainly based on bag-of-words approaches associated with handcrafted rules. One of the most noticeable works of this period was undoubtedly IBM Watson (Ferrucci, 2012) which started to be developed in 2007 and was able to defeat two high-ranked players in a nationally televised two-game Jeopardy.

Starting from 2013, machine learning approaches for question-answering emerged. Multiple datasets with human-labeled question/answers were created. Richardson et al. (2013) introduced MCTest, a question-answering dataset with 660 stories and associated questions with simple baselines based on lexical features. Other line of work focused on large-scale question-answering over knowledge bases (Berant et al., 2014, 2013; Fader et al., 2014)

2015 marked the emergence of deep learning architectures for question-answering. Models such as THE ATTENTIVE READER, THE DEEP LSTM READER, THE IMPATIENT READER where proposed in Hermann et al. (2015) together with a large scale question-answering dataset CNN/Daily Mail. This work showed that these models outperform traditional symbolic-matching models by a considerable margin; 13% of improvement over the best Word distance baseline. Later Chen et al. (2016) showed that a slight modification of this architecture can still improve SOTA by more than 8 points. One of the most popular question-answering datasets of all time was released in 2016. The Stanford Question Answering Dataset (SQUAD) (Rajpurkar et al., 2016) contains more than $107,000$ question/answer pairs about the top 500 Wikipedia articles. It is the first large-scale question-answering dataset with human-generated questions and high-quality labeled

[2] https://trec.nist.gov/data/qa.html
[3] http://www.lrec-conf.org/lrec2002/lrec/wksh/QuestionAnswering.html

answers. It has been a huge breakthrough for the community and has influenced lots of following papers. This dataset has been commonly adopted as a standard baseline to evaluate question-answering models. Lots of following works have successively improved SOTA on this task. Among the most noticeable, we can highlight Match-LSTM (Wang and Jiang, 2017), the Bidirectional Attention Flow reader (BiDAF) (Seo et al., 2017), the R-Net (Wang et al., 2017b), and QANet (Yu et al., 2018b) based on convolutional neural networks.

Finally, during the last couple of years, these architectures for question-answering have been quickly outperformed by BERT-like models (Devlin et al., 2019) as presented in Section 2.1.2. This type of architectures (Devlin et al., 2019; Yang et al., 2019; Clark et al., 2020) nowadays achieves *super-human* performances and occupy most of the top positions on the SQuAD leaderboard [4] and other question-answering benchmarks.

During the last 50 years, all these works have contributed to formally define the question-answering task as a supervised learning problem. The answer can take multiple aspects, and we can divide the question-answering task into four main categories depending on the required type of answer. In the following paragraphs, we formally define these four categories, illustrate them with examples from popular datasets and define the metrics used to evaluate the models.

2.2.1 Cloze style question-answering

Cloze (Taylor, 1953) style of questions can be created by replacing a given word of the query by a placeholder. This is very convenient to generate examples at scale as it does not necessarily require human annotation. Some of the popular datasets with cloze questions being the CNN/DailyMail (Hermann et al., 2015) dataset which consists of cloze questions over news articles, the Children's Book Test (Hill et al., 2016) where the authors use freely available books from the Project Gutenberg[5] to generate the corpus, and WHO DID WHAT (Onishi et al., 2016) which contains cloze questions that require retrieving person named

[4] https://rajpurkar.github.io/SQuAD-explorer/
[5] https://www.gutenberg.org/

Story: (CNN) Sabra Dipping Co. is recalling 30,000 cases of hummus due to possible contamination with Listeria, the FDA said wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on march 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short - term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

Question: a random sample from a @placeholder store tested positive for Listeria monocytogenes
Answer: Michigan

Figure 2.3: Cloze style question-answering example from CNN/DailyMail dataset (Hermann et al., 2015).

entity from LDC English Gigaword newswire corpus [6]. Figure 2.3 presents an example of such type of question.

Evaluation metric: The accuracy score is computed between predicted answers and gold labels to evaluate the systems.

2.2.2 Multi-choice question-answering

This category gathers the datasets where the answer needs to be selected among a given set of candidates. It has been a popular setup with MCTest (Richardson et al., 2013), RACE dataset (Lai et al., 2017) and multiple datasets proposed by the AllenAi's Aristo Project [7]. Figure 2.4 presents an example from the RACE corpus.

Evaluation metric: Similar to cloze style questions, the performance of a system is evaluated with its accuracy.

Story: In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her. When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
[...] The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Question: The first postage stamp was made
Candidates: A) in England B) in America C) by Alice D) in 1910

Figure 2.4: Multi-choice question-answering example from RACE dataset (Lai et al., 2017).

2.2.3 Extractive question-answering

In this category, the answer needs to be extracted from a given input document. It is probably the most used category of questions in the recent applications with datasets such as SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), NewsQA (Trischler et al., 2017) or Natural Questions (Kwiatkowski et al., 2019). An example from the SQuAD dataset is depicted in Figure 2.5.

Evaluation metrics: The performance of a model is evaluated with the two following scores:

- Exact Match (EM), a binary signal that measures whether the answer is correct or not,
- F_1 Score, which computes the average character overlapping between the extracted answer and the gold truth.

[6] https://catalog.ldc.upenn.edu/LDC2003T05
[7] https://allenai.org/aristo

Story: The rainforest contains several species that can pose a hazard. Among the largest predatory creatures are the black caiman, jaguar, cougar, and anaconda. In the river, electric eels can produce an electric shock that can stun or kill, while piranha are known to bite and injure humans. Various species of poison dart frogs secrete lipophilic alkaloid toxins through their flesh. There are also numerous parasites and disease vectors. Vampire bats dwell in the rainforest and can spread the rabies virus. Malaria, yellow fever and Dengue fever can also be contracted in the Amazon region.

Question: What fish living in the Amazon river is known to bit humans?
Answer: piranha

Figure 2.5: Extractive question-answering example from SQuAD dataset (Rajpurkar et al., 2016).

2.2.4 Free-form question-answering

Question: How do Jellyfish function without brains or nervous systems?

Answer: Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They're invertebrate: creatures with-out a backbone. Most jellyfish have really short life spans.Sometimes just a couple of hours. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They're over 95% water.

Supporting documents: [...] Jellyfish do not have brains, and most barely have nervous systems. They have primitive nerve cells that help them orient themselves in the water and sense light and touch. [...] While they don't possess brains, the animals still have neurons that send all sorts of signals throughout their body. [...] They may accomplish this through the assistance of their nerve rings. Jellyfish don't have brains, and that's just where things begin. They don't have many of the body parts that are typical in other animals. [...]

Figure 2.6: Free-form question-answering example from ELI5 dataset (Fan et al., 2019).

The last category corresponds to free-form answers. There is no candidate answer nor pieces of text to extract, but the system needs to generate the answer. Some datasets have been recently proposed such as CoQA (Reddy et al., 2019), NarrativeQA (Kociský et al., 2018) or ELI5 (Fan et al., 2019). This has been a less popular framework for question-answering, especially because the evaluation of the proposed model is more challenging than for other tasks. Figure 2.6 shows an example from the Eli5 dataset.

Evaluation metrics: This task is evaluated with standard metrics from natural language

generation such ROUGE score (Lin, 2004), BLEU score (Papineni et al., 2002) or Meteor (Banerjee and Lavie, 2005).

2.2.5 ReviewQA: a relational aspect-based opinion reading dataset

We believe that evaluating the task of sentiment analysis through the setup of question-answering is a relevant playground for machine-reading research. Indeed natural language questions about the different aspects of targeted venues are typical kinds of questions we want to be able to ask to a system. In this context, we introduce a set of reasoning questions types over the relationships between aspects. We propose ReviewQA, a dataset of natural language questions over hotel reviews. These questions are divided into 8 groups regarding the competency required to be answered. In this section, we describe each task and the process followed to generate this dataset.

Hotel: BEST WESTERN Corona
Title: Convenient Location. Helpful Staff.
Overall rating: ★★★★★

Comment: I just needed a place to sleep and this place was ideally located for my meetings. Plimlico tube is only a few minutes walk. Room was small but clean. Staff very helpful. Breakfast OK.

Ratings			
Service	★★★★★	Location	★★★★★
Rooms	★★★★★	Cleanliness	★★★★★

Figure 2.7: An example from the original dataset.

Original data: We used a set of reviews extracted from the TripAdvisor website and originally proposed in (Wang et al., 2010a) and (Wang et al., 2011). This corpus is available here[8]. Each review comes with the name of the associated hotel, a title, an overall rating, a comment and a list of rated aspects. From 0 to 7 aspects, among *value, room, location,*

[8] http://www.cs.virginia.edu/ hw5x/Data/LARA/TripAdvisor/TripAdvisorJson.tar.bz2

cleanliness, check-in/front desk, service, business service, can possibly be rated in a review. Figure 2.7 displays a review extracted from this dataset.

Relational reasoning competencies: Starting with the original corpus, we aim at building a machine-reading task where natural language questions will challenge the model on its understanding of the reviews. Indeed learning relational reasoning competencies over natural language documents is a major challenge of the current reading models. These original raw data allow us to generate relational questions that can possibly require a global understanding of the comment and reasoning skills to be treated. For example, asking a question like *What is the best aspect rated in this comment ?* is not an easy question that can be answered without a deep understanding of the review. It is necessary to capture all the aspects mentioned in the text, to predict their rating and finally to select the best one. The tasks and the dataset we propose are publicly available here[9].

We introduce a list of 8 different competencies that a reading system should master in order to process reviews and text documents in general. These 8 tasks require different competencies and a different level of understanding of the document to be well answered. For instance, detecting if an aspect is mentioned in a review will require less understanding of the review than explicitly predicting the rating of this aspect. Table 2.1 presents the 8 tasks we have introduced in this dataset with an example of a question that corresponds to each task. We also provide the expected *type* of the answer (Yes/No question, rating question...). It can be an additional tool to analyze the errors of the readers.

We sample 100.000 reviews from the original corpus. We explicitly favor reviews containing an important number of words. On average, a review contains 200 words. Indeed these long reviews are most likely to contain challenging relations between different aspects. A short review which deals with only a few aspects is more likely to not be very relevant to the challenge we want to propose in this dataset. Figure 2.3 displays the distribution of the ratings per aspect in the 100.000 reviews we based our dataset. We can

[9]http://www.europe.naverlabs.com/Blog/ReviewQA-A-novel-relational-aspect-based-opinion-dataset-for-machine-reading

Task id	Description/Comment	Example	Expected answer
1	**Detection of an aspect in a review.** This is a fundamental task. Its objective is to measure how well a model is able to detect whether an aspect is mentioned or not in a review.	Is sleep quality mentioned in this review?	Yes/No
2	**Prediction of the customer general satisfaction.** This second task measures how well a model is able to predict the overall positivity or negativity of a given review.	Is the client satisfied by this hotel?	Yes/No
3	**Prediction of the global trend of an aspect in a given review.** This task measures the satisfaction of a client per aspect. This is a precision over the last task since a client can be globally satisfied by a hotel but not satisfied regarding a certain aspect.	Is the client satisfied with the cleanliness of the hotel?	Yes/No
4	**Prediction of whether the rating of a given aspect is above or under a given value.** This evaluate more precisely how the reader is able to infer the rating of an aspect	Is the rating of location under 4?	Yes/No
5	**Prediction of the exact rating of an aspect in a review.** This task measures precisely the satisfaction of a client regarding an aspect. This is the finest measure that can be extracted from the review.	What is the rating of the aspect Value in this review?	A rating between 1 and 5
6	**Prediction of the list of all the positive/negative aspects mentioned in the review.** To answer a question of this type, the system needs to detect all the aspects that are mentioned in the review and their associated polarity. This question measures the capability of a model to filter positive and negative information.	Can you give me a list of all the positive aspects of this review?	a list of aspects
7.0	**Comparison between aspects.** Depending on the case, this question can require the model to understand precisely the level of satisfaction of the user regarding the two mentioned aspects.	Is the sleep quality better than the service in this hotel?	Yes/No
7.1		Which one of these two aspects, service, location, has the best rating?	an aspect
8	**Prediction of the strengths and weaknesses in a review.** This is probably the hardest task of the dataset. It requires a complete and precise understanding of the review. To perform well on this task, a model should probably master all the previous tasks.	What is the best aspect rated in this comment?	an aspect

Table 2.1: Descriptions and examples of the 8 tasks evaluated in ReviewQA.

	Train	Test	Total
# documents	90.000	10.000	100.000
# queries	528.665	58.827	587.492

Table 2.2: Repartition of the questions into the train and test set.

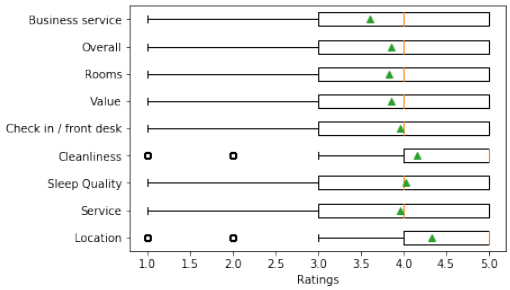

Table 2.3: Distribution of the ratings per aspect.

see that the average values of these ratings tend to be quite high. It could have introduced bias if it was not the case for all the aspects. For example, we do not want that the model learns that in general, the service is rated better than the location and then answer without looking at the document. Since this situation is the same for all the aspects, the relational tasks introduced in this dataset remain relevant.

Then we randomly select 6 tasks for each review (the same task can be selected multiple times) and randomly select a natural language question that corresponds to this task. The process to generate natural language questions is the following: First, for each task, we created a set of patterns corresponding to the evaluated competencies. In these patterns, the aspects have been replaced by placeholders. For instance, *Is @placeholder1 mentioned in this review?*, *Which one of these two aspects @placeholder1, @placeholder2 has the best rating?* are respectively question patterns for task 1 and 7. Second, internally to the NAVER LABS Europe research center, we crowdsourced reformulations of these questions where people were asked to rephrase the patterns while keeping the placeholders. Then we generated the questions of the dataset by replacing the placeholders with several aspects to create coherent questions for each review. Finally, we used a back-translation process to produce additional rephrasing of the questions as it has been suggested in Yu et al. (2018a) to augment the training dataset.

The final dataset we propose is composed of more than 500.000 questions about 100.000 reviews. Table 2.2 shows the repartition of the documents and queries into the train and test set. Each review contains a maximum of 6 questions. Sometimes less when it is not possible to generate all. For example, if only two or three aspects are mentioned in a review, we will be able to generate only a small set of relational questions. A majority of the tasks we introduced, even if they possibly require a high level of understanding of the document and the question, are binary questions. It means that in the generated dataset, the answers *yes* and *no* tend to be more present than the others. To balance in a better way the distribution of the answers, we chose to affect a higher probability of sampling to the tasks 5, 6, 7.1, 8. Indeed, these tasks are not binary questions and required an aspect name as the answer.

2.3 Automatic Text Summarization

In addition to question-answering, human language understanding can also be evaluated through the task of automatic summarization. Automatic Text Summarization refers to the ability of a system to automatically reduce a given piece of text to its essential content. It is a very popular natural language processing task due to its evident utility for numerous applications such as summarizing news articles, legal documents, web content, and so on. The produced summary should be coherent and contain all the essential information from the source text. Interest in this domain started around the 1950s. IBM Auto-Abstracts (Luhn, 1958; Baxendale, 1958) is one of the first works in this domain which aims to create abstracts from scientific and engineering published papers. The task of automatic text summarization can be divided in two main categories, namely **abstractive** and **extractive** summarization.

Extractive Summarization is the process of selecting passages, often sentences, of the source document to construct a summary. A major challenge of this approach is to identify important pieces of the document and combine them together to create a coherent and concise summary without redundancies. This framework has been the most used in early summarizers.

Abstractive summarization is the task of generating a summary after having read the source document. It requires understanding and rephrasing the essential idea of the source document. Compared to extractive summarization, an additional challenge of this type of model is the necessity to generate a linguistically fluent and well-written summary. On the contrary, this is not an issue when the system copies sentences from the source document.

In the 1990s, this task started to attract more interest, and multiple extractive approaches were proposed. Constructing literature abstracts was an important trend at this time (Paice, 1990). Kupiec et al. (1995b) developed a trainable summarization program based on statistical technics on a corpus of 188 documents/summaries. Carbonell and Goldstein (1998) focused on a greedy approach to reorder sentences of a document. Graph-based approaches such as the popular LexRank algorithm (Erkan and Radev, 2004) were also developed along with constraint optimization-based methods such as McDonald (2007).

The modern history of text summarization started around 2015 with neural models. Inspired by sequence-to-sequence (seq2seq) models introduced in Sutskever et al. (2014a) multiple abstractive summarizers emerged. The progress of Bahdanau et al. (2015) with seq2seq models for machine translation were directly adapted to the summarization task. Rush et al. (2015) proposed a seq2seq summarizer based on convolutional models. Chopra et al. (2016); Nallapati et al. (2016b) built on this work and proposed a similar approach using RNNs. At this time, large-scale summarization datasets started to appear. Hu et al. (2015) introduced a Chinese dataset with more than 2 million examples. Later, Cheng and Lapata (2016a) adapted the DeepMind's DailyMail news article dataset (Hermann et al., 2015), presented in last section, to the summarization framework. Together with the dataset, the authors proposed a general framework for document summarization with hierarchical representations and an attention-based extractor. The proposed neural approaches achieve SOTA results without relying on any linguistic annotation.

In addition to abstractive methods, multiple extractive works were proposed. Nallapati et al. (2017) developed SummaRuNNer, a recurrent sequence classifier for extractive summarization. It is one of the early methods to adopt encoders based on RNNs. NeuSUm (Zhou et al., 2018) proposed a joint scoring/selection of sentences to improve results on

this task. Finally, Liu et al. (2019a) introduced Sumo, a structured attention model with dependency tree representations.

Some hybrid approaches combining extractive and abstractive solutions were also developed such as CopyNet (Gu et al., 2016), Forced-Attention Sentence Compression Model (Miao and Blunsom, 2016) or PtGEN (See et al.) which incorporates pointer networks (Vinyals et al., 2015) into an abstractive summarizer.

The last couple of years have seen the Transformer revolution of the NLP with large pre-trained language models. Automatic text summarization has also benefited from these new models. BertSUM (Liu and Lapata, 2019b) introduced an adaptation of the BERT (Devlin et al., 2019) model with classifier heads to label sentences to select for the summary. BART (Lewis et al., 2020) proposed a denoising autoencoder for pre-training seq2seq models. It improve by 6 ROUGE points the current SOTA on the XSum dataset (Narayan et al., 2018) Following this work, Zhang et al. (2020a) proposed PEGASUS, a novel Transformer-based seq2seq model, pre-trained with gap-sentences generation, a novel self-supervised pre-trained objective. It sets a new SOTA for abstractive models on XSum and CNN/DailyMail summarization datasets. In Chapter 5 we propose an approach to use the pre-trained language models for long-document extractive summarization, which remains an issue with recent SOTA extractive models due to the memory complexity of the Transformer.

Evaluation metrics: Manual or semi-manual framework to evaluate the quality of a produced summary have been proposed in Nenkova and Passonneau (2004); Passonneau et al. (2013). However, these are time-consuming and not suitable to evaluate summaries at scale. Automated metrics such as BLEU score (Papineni et al., 2002), ROUGE (Lin, 2004) or Meteor (Banerjee and Lavie, 2005) have been proposed to overcome this issue. Multiple works developed improvements over ROUGE score (Ng and Abrecht, 2015; Ganesan, 2018; ShafieiBavani et al., 2018). However, none of these latest methods still convince the community, and most of the works are currently evaluated with the original ROUGE metric.

Part I

Contributions to Question-Answering

In Chapter 2, we presented the question-answering task as a widely used proxy to evaluate the reading comprehension of a given model. It has been a very active research area during the last years. Many contributions have proposed improvements of existing models and novel architectures until achieving *super-human* performances on certain benchmarks. However, most of the reading methods proposed so far require lots of training data and tend to suffer from a lack of robustness against noisy examples. Besides, current models tend to be good at answering factual questions by detecting similar patterns between question and document but are lacking actual reasoning capabilities. From these observations, we decided to explore two directions described in the following of this chapter. The first one is related to adversarial learning and self-play as a novel approach to train machine comprehension models. We propose an adversarial protocol composed of a couple of models that compete against each other to achieve better performance. The second direction described in this chapter is related to multi-hop reasoning for question-answering. Most of the recent works have been focusing on questions associated to a single, relatively short paragraph. As a consequence, several models tend to rely on advanced matching pattern strategies to detect answers from the documents. On the contrary, in the multi-hop question-answering configuration, a given question cannot be answered from a single paragraph but requires a model to understand multiple pieces of evidence.

The remaining of this first part is divided into two chapters. Chapter 3 describes the proposed adversarial protocol, with associated experiments on three datasets to validate its effectiveness. This work has been published during the thesis in Grail et al. (2018). Chapter 4 describes the multi-hop question-answering task and our proposition, the Latent Question Reformulation Network.

Chapter 3

Adversarial Learning for Text Comprehension

The publication of many large datasets (Rajpurkar et al., 2016; Nguyen et al., 2016; Joshi et al., 2017) has contributed to significant advancement in machine comprehension and question-answering. Recent neural models for machine comprehension now outperform human performance on several of these benchmarks, and there is a lot of novel and promising research on parametric models that feature reasoning capabilities using techniques such as attention and memory. Indeed, as of today, the Human Baseline is rank at position 15 on the GLUE (Wang et al., 2019a) benchmark, composed by the aggregated scores on a collection of 11 various natural language processing tasks. The work in the machine reading field is mainly limited to supervised learning which makes it strictly dependent on the availability of annotated datasets that remains costly to produce. Since the 1990's an increasingly common research activity has been dedicated to self-play and adversariality to overcome this dependency and allow a model to exploit its own decisions to improve itself. Some famous examples are related to policy learning in games. TD-Gammon (Tesauro, 1994) was a neural network controller for backgammon which achieved near top player performance using self-play as learning paradigm. DeepMind's AlphaGo (Silver et al., 2016)

used the same paradigm to win against the currently best human Go player. In the following of these works, AlphaZero (Silver et al., 2017) then MuZero (Schrittwieser et al., 2019) demonstrate the effectivness of self-play when associated to reinforcement algorithms to solve tasks such as chess, shogi or Atari games. The major advantage of such a setting is to alleviate the learning procedure's dependency on an available annotated data. Two models can be set up to learn and improve their performance by acting one against the other in so-called sparring patterns.

Inspired by these learning strategy, we propose to adapt this paradigm of competitive neural networks to machine reading. We developed an approach were two models are in competition to achieve the best performance on a question-answering task while being robust to adversarial perturbations. On the first hand, a **reader network** is trained to learn to answer questions regarding a passage of text. On the other hand, an **obfuscation network** learns to obfuscate words of a given passage in order to minimize the probability of the reading model to successfully answer the question. We developed a sequential learning protocol in order to gradually improve the quality of the models. The proposed idea is to challenge the model by learning to increase the complexity of the task.

This paradigm separates itself from the current approach of joint question and answer learning from text as proposed by (Wang et al., 2017a). Indeed, rather than using question generation as regularizer of a reader model, we suggest using adversarial training to free us from the constraint of strict and bounded supervision and to enhance the robustness of the answering model.

The contributions presented in the following of this section can be summarized as follows: (1) We propose a new learning paradigm for machine comprehension based on adversarial training. (2) With experiments on several machine reading corpora and with several neural architectures, we show that this methodology allows us to overcome the requirement of strict supervision to improve performances of these reading architectures. (3) The attention mechanism allows the visualization of passages considered as meaningful by the obfuscation network.

In the following of this chapter, we discuss the related work in Section 3.1. The proposed adversarial learning protocol and its associated models are described in Section 3.2. In Section 3.4 we validate the our approach on three datasets and present several visualizations of the protocol effectiveness.

3.1 Related Work

3.1.1 Curriculum Learning

In most cases, training neural networks is done by learning from batches of examples randomly selected among the training data. However, we know that humans and animals learn much better when examples are presented in an intelligent way. This strategy allows them to gradually learn more advanced concepts. In this context, curriculum learning has been described in Elman (1993); Rohde and Plaut (1999); Krueger and Dayan (2009) at the frontier of cognitive science and machine learning and show its effectiveness to improve the training process. More recently, Bengio et al. (2009) formalized different curriculum strategies in the context of machine learning and demonstrate its utility on pattern recognition and language modeling tasks with deep neural networks. While most of the work is designed around handcrafted curricula, there have been recent works focusing on automatically building curriculum strategies (Graves et al., 2017; Jiang et al., 2015; Sukhbaatar et al., 2018; Matiisen et al., 2020). The work described in the following of this chapter is closely related to automated curriculum learning. While we do not assume that it is possible to *a priori* classify examples of the dataset by difficulties, we aim at creating novel examples, derived from the original ones, that gradually become harder as the model becomes better. We detail how these harder examples are automatically creating with an adversarial couple of models in the next section.

3.1.2 Adversarial Learning

The idea of using an adversarial learning protocol has been very popular during the last couple of years, particularly in the field of generative models. Indeed Generative Adversarial Networks (GANs), introduced in Goodfellow et al. (2014), have now lots of applications and allowed the training protocol to go beyond the strict supervision of the answer. The main principle of Generative Adversarial Networks (GANs) is to train jointly two adversarial models. These two models are challenging each other with opposing objectives and jointly progressing in the task they are designed for. In machine reading, it has been recently observed that answering a question regarding a text passage and predicting the question regarding a text passage are interesting tasks to model jointly. Consequently, several papers have proposed using the question generation as a regularization task to improve the passage encoding model of a neural reader (Yuan et al., 2017; Wang et al., 2017a). In this work, we acknowledge that these two tasks may indeed be complementary but we believe adversarial training in two player games will lead to similar advantages than those observed previously. As generating a question for a passage is hard, we adapt recent work by Guo et al. (2017) and define the learning of an obfuscation network as a complementary task to the task of learning a reader. Such an obfuscation network tries to find the most meaningful spans of text to obfuscate in a given passage for a given question in order to minimize the probability of the reader successfully answering the question.

3.1.3 Adaptive Dropout

While adversarial examples are a well known research topic in computer vision (Goodfellow et al., 2015), it has not been an active research direction in natural language processing and deep neural network for a long time. Recently, adversarial examples have started to be studied in natural language processing (Miyato et al., 2016; Jia and Liang, 2017; Alzantot et al., 2018; Liang et al., 2018). Jia and Liang (2017) introduces the concept of oversensitivity and demonstrated that a large majority of the recent state-of-the-art deep machine reading models suffer from a lack of robustness regarding adversarial examples. With small

perturbation in the input vectors, they were able to disturb the model completely. In these studies, models suffer from the so-called *catastrophic forgetting*; their average accuracies were decreased by half when tested on corrupted data, i.e., on documents with an additional sentence at the end, which normally should not affect the answer.

One of the attempts to prevent overfitting is to randomly drop network units while training (Srivastava et al., 2014). Such an approach effectively results in combining many different neural networks to make a prediction. In the same spirit, training a model on a dataset with corrupted data is shown to decrease overfitting. Maaten et al. (2013) suggest different ways to corrupt a document, for example by adding noise into the input features; our work refers to what they call the *blankout corruption*, which consist of randomly deleting features in the input documents (texts or images in this case) with probability q. However, learning only from predefined adversarial examples appears sub-optimal since it is not dynamically adapted to the performance of the reader.

In our experiments, we show that randomly corrupted document is not the most efficient way to generate harder examples. We propose to dynamically adapt the the corruption regarding the current the performance of the reader. While obfuscation of one of the keywords can be too hard for the reader at the beginning of the training, obfuscation of a meaningless word is unlikely to have any effect on the reader that is good enough. The learning protocol we propose aims to handle this by training jointly the obfuscation network and the reader in order to adapt the corruption difficulty to the reader's performance.

3.2 Adversarial Reading Protocol

The framework we propose is designed to dynamically generate this adaptive dropout in order to challenge the reader with more and more difficult tasks during the learning stage. We utilize *asymmetric self-play* to train a model called an *obfuscation network* that plays an adversarial game against a *reader*.

The objective of the *obfuscation network* is to iteratively mask several words of the original documents in order to generate new examples that become harder to comprehend for

the *reader*. The obfuscation network is acquiring knowledge about the reader's behaviour during the training, in order to generates increasingly hard adversarial examples. Beyond increasing artificially the size of the available dataset, this adaptive behaviour of the obfuscation network prevents catastrophic forgetting phenomena of the reader. In this section, we first explain our protocol of adversarial training for robust machine comprehension and then describe the reader and obfuscation network models.

The overall framework is a turn-based question-answering game described in Figure 3.1 and algorithm 1. At the beginning of each round t, the obfuscation network masks one word for each document sampled from the training corpus. The ratio of corrupted data / clear data in the dataset is set to $\lambda \in [0, 1]$. Indeed, a too low percentage of corrupted data might not have any effect on the training and a too high one will prevent the reader of learning well. The reader is then trained on a subset of this obfuscated corpus and tested on the remaining subset. Note that both train and test sets contain corrupted data. Finally, the obfuscation network received a set of rewards regarding the reader performance on the obfuscated stories. Given a tuple $(d, d^\dagger, q)$ where d is the original document, $d^\dagger$ the document with an obfuscated word proposed by the obfuscation network and q the associated question, the reward r given to the obfuscation network is defined as follows:

$$r = \begin{cases} 1 & \text{if the reader answers well on } d \text{ and fail on } d^\dagger \\ 0 & \text{otherwise.} \end{cases}$$

The reward given to the obfuscation network is a direct measurement of the impact of the obfuscation on the reader performance. All the previously collected rewards are stored and used for experience replay throughout the turns. After each learning turn, all the parameters of the obfuscation network are reinitialized and retrained on all the recorded rewards. Throughout the turns, the obfuscation network accumulates information about the reader behaviour and proposes more challenging tasks as the game continues. Among the corrupted documents that the obfuscation network proposes to the reader, 80% of the documents maximize the probability of fooling the reader from the obfuscation network

point of view and 20% are randomly corrupted in order to ensure exploration. Finally, the reader keeps improving through time and any catastrophic forgetting is compensated at the next turn of the obfuscation network by focusing on these errors.

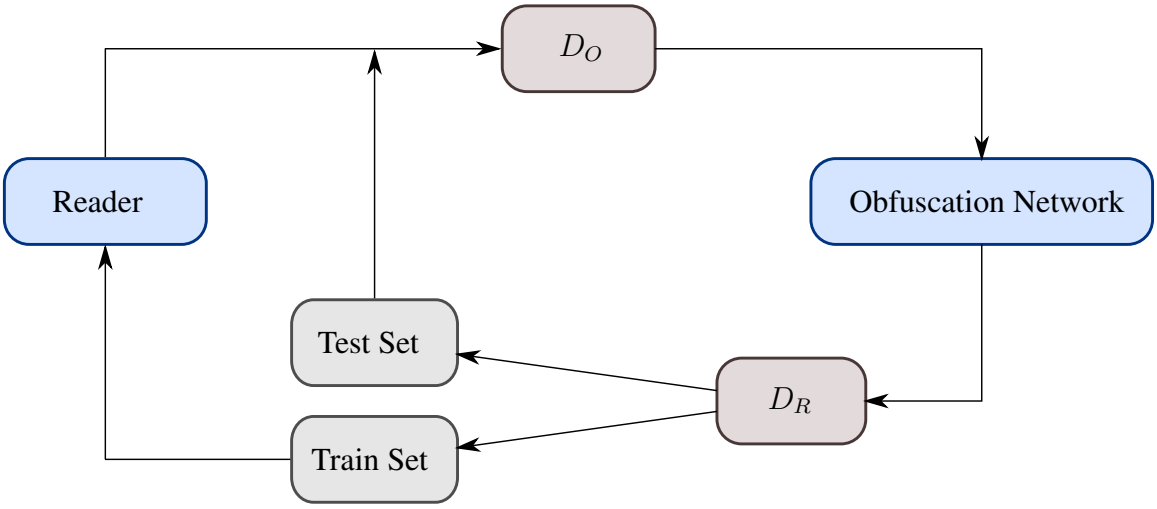

Figure 3.1: Adversarial learning protocol with $D_R = \{d_i, q_i, a_i\}_i$ the reader dataset composed by tuples (document, question, answer) and $D_O = \{d_i, q_i, a_i, r_i\}_i$ the obfuscation network dataset composed by tuples (document, question, answer, reward from the reader).

To more formally specify loss functions for the reader and the obfuscation network, let $\widehat{a}_{ij} \triangleq P(ans_{ij}|q_i, d_i^\dagger)$ denote the reader's predictive probability for ans_{ij} being the correct answer to the question q_i for $j \in [0, n]$ where n is the number of possible answers. Let ij^* be the index of the correct answer. The reader is trained to minimize the cumulative log-loss (cross-entropy) for N questions

$$\mathscr{L}_{\text{Reader}} = -\sum_{i=1}^{N} \log \widehat{a}_{ij^*}. \tag{3.1}$$

The obfuscation network is trained to fool the reader, so it suffers a loss when it fails to predict whether the reader gives a correct answer ans_{ij^*}. By denoting the indicator of the reader answering the question q_i wrong by $fail_i \in \{0, 1\}$ and obfuscation network's estimate of the probability of this failure by $\widehat{a}_i \triangleq P(fail_i = 1|q_i, d_i^\dagger)$, the obfuscation

Algorithm 1: Adversarial training

input: Let I be the initial set of data $\{(d, q, a)\}_i$ where d, q, a are sequences of index representing a document, a question and an answer.
Let A be the training set (80% of I)
Let B be the validation set (10% of I)
Let C be the testing set (10% of I)
Let D be an empty dataset
$t = 0$
while $t <$ NB_MAX_EPOCHS **do**
  Split A into A_1 (80%) and A_2 (20%)
  if t = 0 **then**
    Let $A_1^\dagger$ be A_1 with 20% of random corruption
    Let $A_2^\dagger$ be A_2 with 100% of random corruption
  else
    Reinitialize all the parameters of the obfuscation network
    Train the obfuscation network on D
    Let $A_1^\dagger$ be A_1 with 20% of data corrupted by the obfuscation network
    Let $A_2^\dagger$ be A_2 with 100% of data corrupted by the obfuscation network
  end if
  Train one epoch of the reader on $A_1^\dagger$
  for all $((d, q, a) \in A_2, (d^\dagger, q, a) \in A_2^\dagger)$ **do**
    Let r be the reward given to the obfuscation network
    if the reader succeed on d and fails on $d^\dagger$ **then**
      $D \leftarrow \{D \cup (d^\dagger, q, a, r = 1)\}$
    else if the reader succeed on d and succeed on $d^\dagger$ **then**
      $D \leftarrow \{D \cup (d^\dagger, q, a, r = 0)\}$
    end if
  end for
  Let ε_t be the empirical error of the reader on B
  if $\varepsilon_t > \varepsilon_{t-1}$ **then**
    Stop the learning
  end if
  t $\leftarrow$ t $+ 1$
end while
Report the empirical error of the reader on C

network's loss is defined as

$$\mathscr{L}_{\text{ObfNet}} = -\sum_{i=1}^{N} fail_i \log \widehat{a}_i + (1 - fail_i) \log(1 - \widehat{a}_i). \tag{3.2}$$

3.2.1 Reader network

To illustrate this work, we investigate two types of neural architectures: a memory based architecture with a Gated End-to-End Memory Network (Liu and Perez, 2017) (`GMemN2N`) and a multi-layer attention based architecture largely inspired by the recent `R-Net` (Wang et al., 2017b) excepted for its output layer, adapted to the format of the datasets used in this work. These two architectures are competitive models for machine reading and most of the recent models are a combination of layers included in these two architectures. Paragraphs below describe these two architectures and how we have integrated them in the adversarial learning protocol.

3.2.1.1 Gated End-to-End Memory Network reader

The first model used as a reader is a Gated End-to-End Memory Network Liu and Perez (2017), `GMemN2N` (Figure 3.2). This architecture is based on two different memory cells and an output prediction. An input memory representation $\{m_i\}$ and an output representation $\{c_i\}$ are used to store embedding representations of inputs. Suppose that an input of the model is a tuple (d, q) where d is a document, i.e., a set of sentences $\{s_i\}$, and q is a query about d. The entire set of sentences is converted into input memory vectors $m_i = A\Phi(s_i)$ and output memory vectors $c_i = C\Phi(s_i)$ by using two embedding matrices A and C. The question q is also embedded using a third matrix B, $u = B\Psi(q)$ of the same dimension as A and C, where Φ and Ψ are respectively the document embedding function and the question embedding function described in the next paragraph. The input memory is used to compute the relevance of each sentence in its context regarding the question, by computing the inner product of the input memory sentence representation with the query. A $\mathrm{softmax}$ is then used to map the inner product to a probability. The response $o = \sum_i p_i c_i$

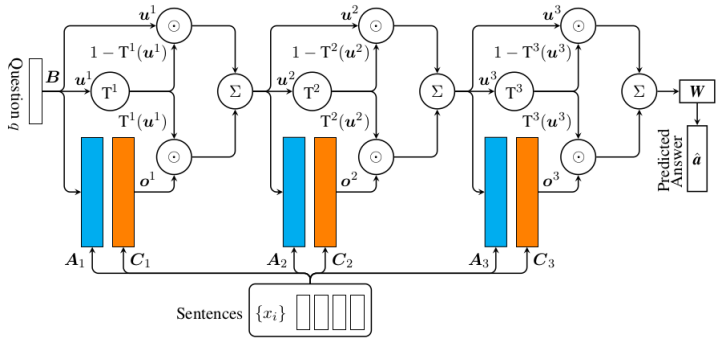

Figure 3.2: Gated End-to-End Memory Network Liu and Perez (2017).

from the output memory is the sum of the output memory vectors $\{c_i\}$ weighted with the sentence relevancies calculated before $p_i = \mathrm{softmax}(\boldsymbol{u}^T m_i)$. A gated mechanism is used when we update the value of the controller u:

$$T^k(u^k) = \sigma(W_T^k u^k + b_T^k), \tag{3.3}$$

$$\mathbf{u^{k+1}} = o^k \odot T^k(u^k) + u^k \odot (1 - T^k(u^k)), \tag{3.4}$$

where W_T^k are matrices of size $d \times d$ and b_T^k a vector of size d with d the size of the memory cells.

Assuming we use a model with K hops of memory, the final prediction is:

$$\widehat{a} = \mathrm{softmax}(W(o^K + u^K) + b), \tag{3.5}$$

where W is a matrix of size $d \times v$ and b a vector of size d with v the number of candidate answers. In this model, we do not use the adjacent or layer-wise weight tying scheme and all the matrix A^k and B^k of the multiple hops are different.

Text and question representations (Figure 3.3): To build the sentence representations, we use a 1-dimensional Convolutional Neural Network (CNN) with a list of filter sizes over all the sentences as proposed in Kim (2014). Let $[s_1, \ldots, s_N]$ be the vectorial

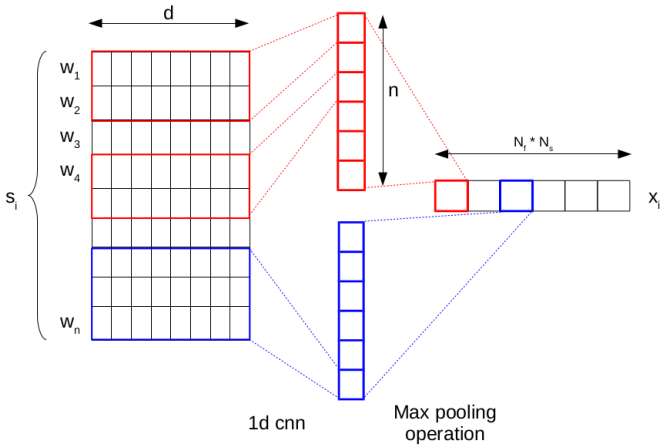

Figure 3.3: An encoded sentence where d is the word embedding size, N_f the number of filters of each size and N_s the number of different filter sizes used.

representation of a document with N sentences where $s_i = [w_{i,1}, w_{i,2}, \ldots, w_{i,n}]$ is the $i-th$ sentence which contains n words. Given a convolutional filter $F \in \mathbb{R}^{h \times d}$ where h is the width of the convolutional window, i.e, the number words it overlaps, the convolutional layer produces:

$$c_{i,j} = f(F \odot [Ew_{i,j}, \ldots, Ew_{i,j+h}]), \forall j \in [1, n-j], \tag{3.6}$$

where $\odot$ is the element-wise multiplication, f a rectified linear unit (ReLU) and E is the embedding matrix of size $d \times V$ where V is the vocabulary size and d the word embedding size. Then, a max pooling operator is applied to this vector to extract features. Given a filter F, after a convolutional operation and a max pooling operation, we obtain a feature $\hat{c}_i = \max_j(c_{i,j})$ from the i^{th} sentence of the text. Multiple filters with varying sizes are used. Assume that our model uses N_s different filter sizes and N_f for each size, we are able to extract $N_s \times N_f$ features for one sentence. The final representation of the sentence is the concatenation of the extracted features from all the filters:

$$\Phi(s_i) = [\hat{c}_{iF_1}, \hat{c}_{iF_2}, \ldots, \hat{c}_{iF_{N_s * N_f}}]. \tag{3.7}$$

Compared to an LSTM encoding the CNN layer is faster and gives better results on the different tasks we evaluated our model. This result seems coherent with recent results of Dauphin et al. (2017). We use a bidirectional GRU to encode the question. The question representation $\Psi(q)$ is the concatenation of the final states of the forward and backward GRU on this question.

3.2.1.2 R-Net based network

The second architecture investigated in this article is based on the state-of-the-art `R-Net` model (Wang et al., 2017b). The main part of the architecture remains the same as the original model, except for the last layer. We replaced the pointer network, originally used to select in the document the span of text that corresponds to the answer, by a fully connected layer followed by a $\mathrm{softmax}$ to output the probability of each candidate word to be the answer. The following lines describe the structure of this architecture, composed of multiple stacked layers.

Encoding layer: Each sentence is tokenized by word and each token is represented by the concatenation of the word level, and character level embeddings. The word level embedding is computed via a lookup table initialized with `GloVe` pre-trained embeddings and the character embedding of a token is the final state of a GRU network over the sequence of its characters. Finally, these tokens are fed to a Recurrent Neural Network (RNN) and the document and question are represented by the intermediate states of this RNN.

Gated question/document attention: Assuming that $d = \{u_i^d\}_{i=0}^N$ and $q = \{u_i^q\}_{i=0}^N$ are the sequences of embedding tokens of the document and the question after the encoding layer with N the length of the document and n the length of the question. Then we compute an attention between the representation of the question and each token of the document. The document is transformed to $d = \{v_i^d\}_{i=0}^N$ with:

$$v_i^d = RNN(v_{i-1}^d, [u_i^d, c_i]),$$

where c_i is an attention vector of the question over the token i of the document. This layer produces a question-aware representation of the document.

Self-attention: So far each token contains information from the question due to the question/document attention layer and from its surrounding context due to the RNN at the end of the encoding part but does not handle long-term dependencies inside the document. The self-attention layer produces an attention between the whole document and each individual token of it. $d = \{h_i^d\}_{i=0}^N$ with:

$$h_i^d = BiRNN(h_{i-1}^d, [v_i^d, c_i]),$$

where c_i is an attention vector of the whole document over the token i.

Output layer: The decision support is the concatenation of the h_i, for $i \in [0, N]$. $o^d =$ concat$(\{h_i\}_{i=0}^N)$ and finally:

$$\widehat{a} = \text{softmax}(Wo^d + b),$$

where W is a matrix of size $N * d \times v$ and b a vector of size d with v the number of candidate answers.

3.2.2 Obfuscation network

The objective of this model is to predict the probability of the reader to successfully respond to a question about a document with an obfuscated word. This estimate will be used by the obfuscation network to determine the position of the obfuscated word in the document which maximizes the probability of the reader to fail its task. We use a similar architecture as the reader, i.e a `GMemN2N` when the reader is a `GMemN2N` and a `R-Net` when the reader is a `R-Net`. However, on the last layer, a sigmoid function is used to predict the probability of the reader to fail on this input: Assuming that o is the decision support of the

obfuscation network, then:

$$\widehat{a} = \sigma(Wo + b), \tag{3.8}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\widehat{a} \in [0, 1]$ is the predicted probability of failure of the reader and W a matrix of size $d \times 1$. We impose this symmetry between the architecture of the reader and of the obfuscation network in order to keep a fair challenge between the two adversary networks.

An input to the reader is a tuple $(d^{\dagger}, q)$ where $d^{\dagger}$ is a document with an obfuscated word. To obfuscate a word, we replace it by the word *unk* for *unknown*. The output of the obfuscation network is a real number $r \in [0, 1]$ which is the expected probability of the reader to fail on the question. The objective of the obfuscation network is to select the corrupted document which maximizes this reward. We use the same text passage and query representation as for the reader, based on a CNN with different filter sizes for the document and the two last hidden states of a bidirectional Gated Rectified Unit (GRU) recurrent network for the question encoding for the `GMemN2N` and based on character and word level embeddings for the `R-Net`. Both models are fully differentiable.

3.3 Baseline Protocol

In addition to the adversarial protocol, we propose a baseline version of it. In this setup, the corruption is made by randomly obfuscating a word in several documents. This is a naive variation of the first protocol where the obfuscation network does not learn from the reader feedback at all. In fact, this protocol is similar to a dropout regularization on the embeddings layer that allows avoiding overfitting the training set. However, the obfuscation is independent of the reader performance; especially, it does not take into account the difficulty of the questions. This learning protocol has strong similarities with the one proposed by Maaten et al. (2013).

3.4 Experiments and Analysis

3.4.1 Datasets

Cambridge Dialogs: The transactional dialog corpus proposed by Wen et al. (2017) has been produced by a crowdsourced version of the Wizard-of-Oz paradigm. It was originally designed for dialog state tracking, but Liu and Perez (2017) have shown that this task could also be considered as a reading task. In such setting, the informable slots provided as metadata to each dialog were used to produce questions for a dialog comprehension task. The dataset deals with an agent assisting a user to find a restaurant in Cambridge, UK. To propose the best matching restaurant, the system needs to extract 3 constraints which correspond to the informable slots in the dialog state tracking task: *Food, Price range, Area*. Given a dialog between an agent and a user, these informable slots become questions for the model we propose. The dataset contains 680 different dialogs about 99 different restaurants. We preprocess the dataset to transform it into a question-answering dataset by using the three informable slot types as questions about a given dialog. After this preprocessing operation, we end up with our question-answering formatted dataset which contains 1,352 possible answers.

Document: *I want the phone number of a moderately priced restaurant with Spanish food.* *La Tasca would fit the bill. Its phone number is 01223 464630.* *Can you tell me what area of town it is located?* *La Tasca is located in the center part of town.* *Thank you, goodbye.* *You're welcome.* **Question:** *What is the area?* **Answer:** *Center.*

Table 3.1: An example from the Cambridge dataset formatted for question-answering task.

TripAdvisor aspect-based sentiment analysis: This dataset contains a total of 235K detailed reviews extracted from the TripAdvisor website and originally released by Wang et al. (2010b). These reviews represent around 1,850 hotels. Each review is associated to

an overall rating, between 0 and 5 stars. Furthermore, 7 aspects: *value, room, location, cleanliness, checkin/front desk, service,* and *business service* are available. We transform the dataset into a question-answering task over a given review. Concretely, for each review, a question is an aspect and we use the number of stars as the answer. This kind of machine-reading approach to sentiment analysis was previously proposed in Tang et al. (2016).

Document: *Service was ok, staff helpful, room was basic, marks on bedding top cover looked like blood, sheets clean, bathroom not so nice, broken tiles on floor, shower head was disgusting and needed to be replaced, location was good, close to the metro and the Colosseum, both only a 10 min walk, liked that the hotel was close to many cafe's restaurant's, disliked the shower in room.*

Question: *How is the cleanliness?*
Answer: 2/5.

Question: *How is the service?*
Answer: 3/5.

Table 3.2: An example from the TripAdvisor dataset.

Children's Book Test (CBT): The dataset is built from freely available books (Hill et al., 2016) produced by Project Gutenberg[1]. The training data consists of tuples (S, q, C, a) where S is the *context* composed of 20 consecutive sentences from the book, q is the *query*, C a set of 10 *candidate answers* and a the *answer*. The query q is the 21^{st} sentence, i.e., the sentence that directly follows the 20 sentences of the *context* and where one word is removed and replaced by a missing word symbol. Questions are grouped into 4 distinct categories depending of the type of the removed word: Named Entities (NE), (Common) Nouns (CN), Verbs (V) and Prepositions (P). This division of answers according to the type of the word that has been removed give a way to evaluate the performance of a model in different situations. It provides relevant information on the strengths and weaknesses of a given architecture. The training contains 669,343 inputs (context+query) and we evaluated our models on the provided test set which contains 10,000 inputs, 2,500 per category. This

[1] https://www.gutenberg.org.

dataset evaluates the capability that a model has to predict a word based on its context.

Document:
1 When she got home she shut herself up in her room and cried.
2 There was nothing for her to do but resign, she thought dismally.
3 On the following Saturday Esther went for an afternoon walk, carrying her Kodak with her.
4 It was a brilliantly fine autumn day, and woods and fields were basking in a mellow haze.
19 Bob and Alf Cropper were up among the boughs picking the plums.
20 On the ground beneath them stood their father with a basket of fruit in his hand.

Question: *21 Mr. Cropper looked at the **XXXXX** and from it to Esther.*
Answer: *proof*
Candidates: *Saturday — boughs — face — father — home — nothing — proof — remarks — smile — woods*

Table 3.3: An example from the CBT dataset.

In this section, we present our experimental settings and the results of this adversarial training protocol on the three datasets presented in Section 3.4.1.

3.4.2 Training details

10% of the dataset was randomly held-out to create a test set. We split the dataset before all the training operations and each protocol was tested on the same test dataset. For the training phase, we split the training dataset to extract a validation set to perform early stopping. We used Adam optimizer (Kingma and Ba, 2015) with a starting learning rate of 0.0005. We set the dropout to 0.9 which means that during training, randomly selected 10% of the parameters are not used during the forward pass and not updated during the backward propagation of error. We also added the gated memory mechanism (Liu and Perez, 2017) that dynamically regulates the access to the memory blocks. This mechanism had a very positive effect on the overall performance of our models. All weights were initialized randomly from a Gaussian distribution with zero mean and a standard deviation of 0.1. We augmented the loss with the sum of squares of the model parameters.

The hyperparameters have been chosen via cross-validation on the validation set of the different datasets. We set the batch size to 16 inputs and we used word embeddings of size 300. We initialized all the embedding matrices with pre-trained `GloVe` word vectors (Pennington et al., 2014) and used random vectors for the words not present in the `GloVe` . It seems that for our experiments CNN encoding does not improve only the overall accuracy of the model compared to LSTM but also the stability by decreasing the variance of the results. So, in practice, we used 128 filters of size 2, 3, 5 and 8 resulting in a total of 512 filters for the one-dimensional convolutional layer.

We repeated each training 10 times for the first two datasets and report maximum and average accuracy. The average value corresponds to the average score over the 10 runs on the test set. Maximum value corresponds to the score on the test set achieved by the model that performed best on the validation set. During the adversarial learning, the dataset contained 70% of clear dialogs and 30% of corrupted dialogs, $\lambda = 0.3$. Inside these corrupted data, 20% were randomly obfuscated by the obfuscation network in order to make it learn from exploration and the obfuscation network maximized its reward for the remaining 80%. Due to the format of the dataset, we slightly modified the output layer of our reader for the CBT task. Instead of projecting on a set of candidate answers, the last layer of the reader made a projection on the entire vocabulary $\widehat{\boldsymbol{a}} = \sigma(M \odot W(o^K + u^K))$ where W is a matrix of size $V \times d$ with V the vocabulary size, $\odot$ the elementwise product and M the mask vector of size V containing 1 if the corresponding word is proposed in the candidate answers, and 0 otherwise.

3.4.3 Results

	Log Reg	ASR	`GMemN2N`			uniform `GMemN2N`			adversarial `GMemN2N`		
hops	×	×	4	5	6	4	5	6	4	5	6
Max	58.4	40.8	82.1	85.8	80.6	85.1	85.8	82.8	82.8	79.8	**88.1**
Mean	58.2	39.5	76.9	74.8	74.2	77.4	77.7	74.9	**79.8**	77.8	79.6

	R-Net	uniform R-Net	adversarial R-Net
Max	88.1	89.5	**90.8**
Mean	87.5	89.2	**90.0**

Table 3.4: Average and maximum accuracy (%) on the Cambridge dataset on 10 replications. In bold, the best result per architecture.

	Log Reg	ASR	GMemN2N			uniform GMemN2N			adversarial GMemN2N		
hops	×	×	4	5	6	4	5	6	4	5	6
Max	59.4	45.2	62.3	62.4	60.5	63.1	61.4	63.1	**64.6**	63.5	62.3
Mean	59.0	42.3	60.8	60.6	58.5	62.3	60.3	59.6	**62.8**	61.2	60.8

	R-Net	uniform R-Net	adversarial R-Net
Max	62.3	63.8	**64.5**
Mean	61.9	62.2	**63.0**

Table 3.5: Average and maximum accuracy (%) on the TripAdvisor dataset on 10 replications. In bold, the best result per architecture.

In this section, we report the results of our implementation of two baselines: a simple logistic regression and an Attention-Sum Reader (Kadlec et al., 2016). Then we present the results of our implementation of the two neural architectures presented in Section 3.2.1, trained with the standard training, the *uniform* training, which is the reader trained with the baseline protocol 3.3 and with our adversarial learning protocol 3.1.

Tables 3.4 and 3.5 display the scores obtained by these models on the Cambridge and TripAdvisor datasets. Each experiment was run 10 times and we report in this table the maximum score on the test set (based on the validation set) and the average score. The precise number of hops needed to achieve the best performance with the GMemN2N is not obvious, so we present all the results for readers and obfuscation networks between 4 and 6 hops.

We observe that the **adversarial learning protocol improves the accuracy** of the

GMemN2N and R-Net compared to the standard and uniform training protocol for all the experiments.

We improve the score of the reader by 2.3 points on the Cambridge task for a GMemN2N with 6 hops compared to the standard training. This adversarial protocol, applied to the R-Net architecture, improves the average score by 2.5 points on this dataset.

The best performance on the TripAdvisor dataset was achieved by the adversarial R-Net. On 10 replications of the experiment, the average accuracy of this model was improved by 1.1 points compared to the standard approach.

The GMemN2N with 4 hops achieved the best performance of this architecture. The accuracy was improved by 1.5 points when the model was trained with our adversarial protocol.

The uniform protocol improves the stability of the performance compared to a standard reader but further improvements were obtained with the adversarial protocol which improved both the overall accuracy and the stability of the performance. Indeed the variance of the results decreased when the training was done with the adversarial protocol, especially for the GMemN2N . Such architecture does not always converge to the optimal minima and the adversarial learning, acting as an adaptive dropout, seems to help the model to generalized better. It is not clear, for this task, whether the number of hops, between 4 and 6, affects the general behaviour, but we achieved the best performance with our adversarial protocol and a reader with 6 hops.

	Log Reg				ASR			
Task	P	V	NE	CN	P	V	NE	CN
Max	56.3	37.1	26.5	25.6	24.7	32.7	22.1	18.3

	GMemN2N				uniform GMemN2N				adversarial GMemN2N			
Task	P	V	NE	CN	P	V	NE	CN	P	V	NE	CN
Max	56.0	58.5	31.9	39.0	58.1	53.6	31.6	34.0	**71.1**	**60.4**	**35.3**	**39.4**

	R-Net				uniform R-Net				adversarial R-Net			
Task	P	V	NE	CN	P	V	NE	CN	P	V	NE	CN
Max	55.0	68.3	44.0	42.6	56.3	68.9	43.8	40.7	**60.0**	**70.0**	**44.5**	**42.9**

Table 3.6: Accuracy (%) on the CBT dataset. In bold, the best result per architecture.

Performance on the CBT dataset are displayed in Table 3.6. Because of the size of this dataset, we didn't repeat the training 10 times but only once. Results of the uniform training seem similar to the performance of the standard reader in this case but **the accuracy of the models trained with our adversarial protocol remains higher than others'**. This last experiment shows that augmenting the data distribution with random adversarial examples might not help the training as it was explained by Jia and Liang (2017). But we show that even in that case the protocol we propose is able to generate smart adversarial examples that will finally help the reader to improve its overall performance.

3.4.4 Visualizations and analysis

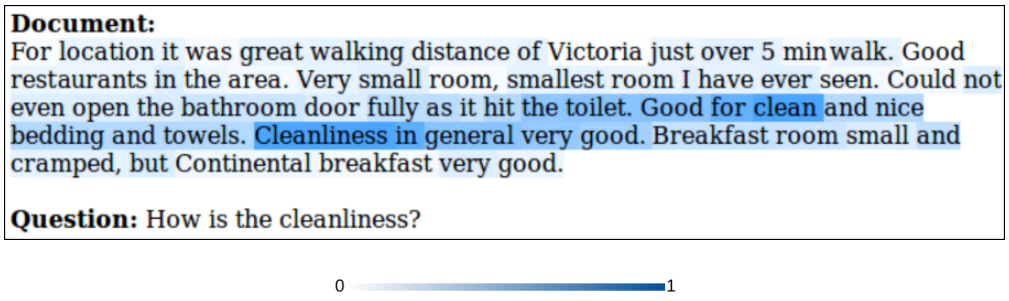

Figure 3.4: Rewards expected by the obfuscation network after 100 rounds over a TripAdvisor review.

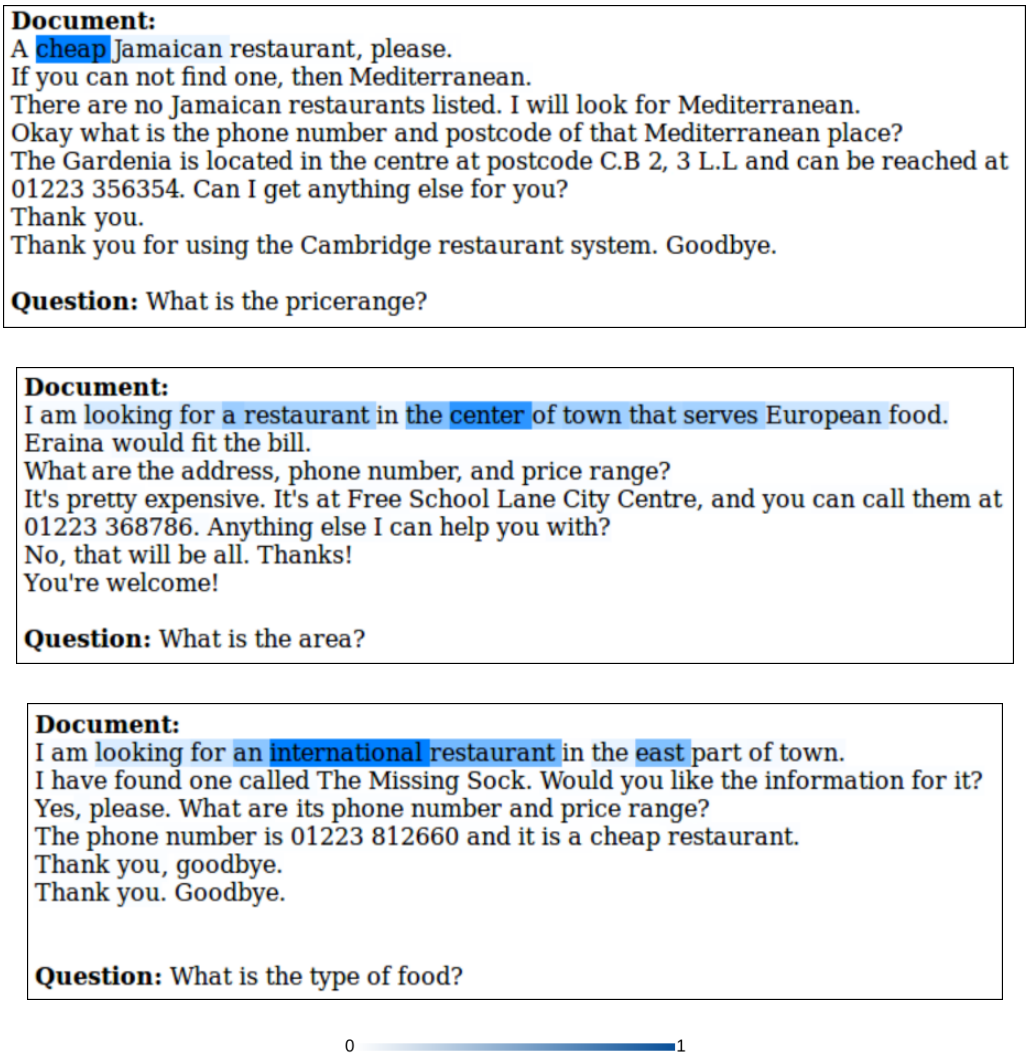

Figure 3.5: Rewards expected by the obfuscation network after 100 rounds over a Cambridge dialog.

In this section, we present a series of analysis of the behavior of the competitive networks to better understand how the adversarial setting affects the training. We propose to analyze the probabilities of obfuscation of the different words of a given input (d, q, a).

In order to better understand how the obfuscation network learns from the reader behaviour during the adversarial protocol, Figure 3.5 depicts the rewards that the obfuscation network expects for each word of a document after several rounds of the game. Given a

tuple (d, q) where d is a clear document and q a query, and assuming the document contains k words, we generate k corrupted documents where one word is obfuscated in each of them. We then feed the obfuscation network with these corrupted data and report the results. The expected rewards from the reader are displayed in green on the document. A strong intensity means that a high reward is expected.

We see that the obfuscation network tends to obfuscate some important keywords of the dialogs in Figure 3.5. Furthermore, the obfuscation network is not pointing on a single word but it points on a word and on its neighborhood. This could be a consequence of the encoding which is not only a representation of a single word but a representation of a word in its context. In Figure 3.4, we can see that the obfuscation network tends to affect a high probability of getting a reward for multiple words of the review. This can be a consequence of the performance of the reader on this dataset. Indeed if the reader is not generally confident about its answers, small changes in the reviews could lead to fool it. However, we can see on the figure that the most probable regions obfuscated by the obfuscation network refer to the cleanliness of the hotel which is coherent with the question.

3.5 Discussion on Masked Language Modeling

In this chapter, we proposed an adversarial learning protocol to train coupled deep neural networks for a question-answering task. We proposed two baselines, a Logistic Regression and an Attention-Sum Reader, on the three datasets used in the experiments. We experimented our adversarial learning protocol on two main types of neural architectures based on state-of-the-art machine reading models at this time: a `GMemN2N` and a `R-Net`. In addition, we compared our adversarial protocol to a protocol based on a uniform corruption of data. On all the reported experiments, the models trained with our novel protocol outperform the equivalent models trained with a standard supervised protocol or a protocol that introduces a uniform noise in the data, which corresponds to the more classic approach of dropout.

We proposed this adversarial learning protocol (Grail et al., 2018) before the introduction of BERT (Devlin et al., 2019) and the emergence of the now popular bidirectional masked-language modeling task. However, these two approaches are closely related, and we aim to discuss their similarities and differences in the following paragraph.

Similar to BERT, our approach emphasizes the fact that word encoding should not depend on a single token but are much more efficient when contextualized with both left and right context of this token. To ensure that the model builds contextualized token representations, we used a similar approach to BERT by obfuscating words of the input sequence. Masking a word constrains the model to use the context of this given word to build a meaningful representation of it. In this work, we used a `GMemN2N` and a `R-Net` with CNN and bidirectional LSTM encoding functions. By design, these two encoding functions allow the model to use both left and right contexts of a masked word to construct its representation.

The major difference between this work and BERT-like approaches concerns the objective of this masking strategy. Indeed, while BERT-like approaches aim to build general representations of words useful for a collection of downstream tasks, we did not design this protocol with a pre-training objective. Compared to BERT-like model, our protocol does not rely on external data and is trained only on the task-related dataset. We directly use this word masking strategy to improve the performance of a given model on a specific final task. As a second difference, we do not randomly mask the words but use an adversarial strategy to optimize the selection of the words to corrupt. Because the goal is not to produce a universal representation of words, we can optimize the masking strategy according to the task objective. In our case, the masking objective is designed to corrupt the most essential words regarding to the reader model. As shown in Figure 3.5 after several steps, the obfuscation network converges to parts of the text that are crucial for the downstream task. This avoids masking words that are useless for the task and forces the model to build better representations of the important ones from their context.

Chapter 4

Multi-hop Machine Reading

We have seen that the ability to automatically extract relevant information from large text corpora remains a major challenge, and how the training can benefit from adversarial learning and self-play. However, most of the current datasets for question-answering focus on the ability to read and extract information from a single piece of text, often composed of few sentences (Rajpurkar et al., 2016; Nguyen et al., 2016). As a consequence, Sugawara et al. (2018) shows that this setup has encourage annotators to produce *easy questions* and influenced the recent state-of-the-art models to be good at detecting patterns and named entities (Devlin et al., 2019; Yu et al., 2018a; Wang et al., 2017b). However, beyond *word matching*, reasoning capabilities are not clearly challenged in these configurations.

Easy questions from Sugawara et al. (2018) are the ones that can be answered with one of the two following heuristics: (1) there is only one possible answer based on expression such as *when, where, how many* present in the question and (2), the answer is in the paragraph's sentence that is the most similar to the question in term of edit distance.

Thus, we decided investigate the multi-hop question-answering task aims at challenging reasoning capabilities of a reader. Multi-hop question-answering requires machine comprehension models to *gather* and *compose* over different pieces of evidence spread across multiple paragraphs. To do so, we propose an original neural architecture that repeatedly reads from a set of paragraphs to aggregate and reformulate information. Besides

sequential reading, our model is designed to collect pieces of information in parallel and to aggregate them in its last layer. Throughout the model, the important pieces of the document are highlighted by what we call a **reading module** and integrated into a representation of the question via our **reformulation module**. The contributions of this section can be summarized as follows:

- We propose a machine reading architecture, composed of multiple token-level attention modules, that collect information sequentially and in parallel across a document to answer a question,
- We propose to use an input-length invariant question representation updated via a dynamic max-pooling layer that compacts information form a variable-length text sequence into a fixed size matrix,
- We introduce an extractive reading-based attention mechanism that computes the attention vector from the output layer of a generic extractive machine reading model,
- We illustrate the advantages of our model on the HOTPOTQA dataset (Yang et al., 2018).

The remainder of the section is organized as follows: first, we present the multi-hop question-answering task and analyses the required reasoning competencie; the related word is discussed in Section 4.1. In Section 4.2, we present our novel reading architecture and present its different building blocks. Section 5.3 presents the conducted experiments, several ablation studies, and qualitative analysis of the results.

The task of extractive machine reading can be summarized as follows: given a document D and a question Q, the goal is to extract the span of the document that answers the question. In this work, we consider the explainable multi-hop reasoning task described in Yang et al. (2018) and its associated dataset: HOTPOTQA . We focus our experiments on the "**distractor**" configuration of the dataset. In this task, the input document D is not a single paragraph but a set of ten paragraphs coming from different English Wikipedia

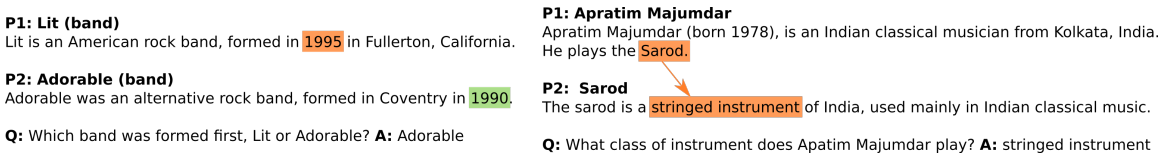

Figure 4.1: Examples of reasoning paths to answer two questions of the HOTPOTQA dataset. In this picture, we do not display the full paragraphs, but only the supporting facts.

articles. Answering each question requires gathering and integrating information from exactly two paragraphs; the eight others are distractors selected among the results of a tf-idf retriever (Chen et al., 2017). These required paragraphs are called the *gold* paragraphs. There are two types of questions proposed in this dataset: *extractive* ones where the answer is a span of text extracted from the document and binary *yes/no* questions. In addition to the answer, it is required to predict the sentences, also called *supporting facts*, that are necessary to produce the correct answer. This task can be decomposed in three subtasks: (1) categorize the answer among the three following classes: *yes, no, text span*, (2) if it is a span, predict the start and end positions of this span in the document, and (3) predict the supporting sentences required to answer the question. In addition to the "**distractor**" experiments, we show how our proposed approach can be used for open-domain question answering and evaluate the entire reading pipeline on the "**fullwiki**" configuration of the HotpotQA dataset. In this configuration, no supporting documents are provided, and it is required to answer the question from the entire Wikipedia corpus.

Among the competencies that multi-hop machine reading requires, we identify two major reasoning capabilities that human readers naturally exploit to answer these questions: sequential reasoning and parallel reasoning. **Sequential reasoning** requires reading a document, seeking a piece of information, then reformulating the question and finally extracting the correct answer. This is called multi-hop question-answering and refers to the *bridge* questions in HOTPOTQA . Another reasoning pattern is **parallel reasoning**, required to collect pieces of evidence for comparisons or question that required checking multiple properties in the documents. Figure 4.1 presents two examples from HOTPOTQA

that illustrate such required competencies. We hypothesize that these two major reasoning patterns should condition the design of the proposed neural architectures to avoid restricting the model to one or the other reasoning skill.

4.1 Related Work

Multi-hop dataset: QAngaroo (Welbl et al., 2018) is another dataset designed to evaluate multi-hop reading architectures. It requires sequentially gather information over documents to answer a question. In this dataset, each question comes with an associated set of candidate answers. Each document has been generated using a knowledge base and the question takes the form of an {subject, predicat} couple where the answer needs to be selected among a predefined list of candidates. The state-of-the-art architectures on this task (Zhong et al., 2019b; Cao et al., 2019) tend to exploit the structure of the dataset by using the candidate spans as an input of the model. In practice, most of the approaches handcraft a graph of the candidate with a selected set of relationship and use Graph Convolution Network to compute the likelihood of each candidate node with respect to the question. In the case of HotpotQA, text and questions have been extracted from Wikipedia and no candidate are given. Furthermore, we observe than only 50 % of the answer can be detected by a Named Entity Recognizer, which could have been a way to extract candidate answers. Other approaches propose to reconstruct the pseudo gold reasoning chain and to use this information as a supervision signal. These facts can explain why successful approaches on QAngoroo have been transfer into the HOTPOTQA dataset.

Multi-hop Machine Comprehension: The question-answering task has recently increased its popularity as a way to assess machine reading comprehension capabilities. The emergence of large scale datasets such as CNN/Daily Mail, (Hermann et al., 2015), SQuAD (Rajpurkar et al., 2016) or MSMARCO (Nguyen et al., 2016) have encouraged the development of multiple machine reading models (Devlin et al., 2019; Wang et al., 2018; Tan et al., 2017). These models are mainly composed of multiple attention layers that update

the representation of the document conditioned by a representation of the question.

However, most of this work focuses on the ability to answer questions from a single paragraph, often limited to a few sentences. Weston et al. (2016); Joshi et al. (2017) were the first attempts to introduce the task of multi-documents question-answering. QAngaroo (Welbl et al., 2018) is another dataset designed to evaluate multi-hop reading architectures. However, state-of-the-art architectures on this task (Zhong et al., 2019b; Cao et al., 2019) tend to exploit the structure of the dataset by using the proposed candidate spans as an input of the model.

Recently, different approaches have been developed for HOTPOTQA focusing on the multiple challenges of the dataset. Nishida et al. (2019) focuses on the evidence extraction task and highlight its similarity with the extractive summarization task. Related works also focus on the interpretation of the reasoning chain with an explicit decomposition of the question (Min et al., 2019b) or a decomposition of the reasoning steps (Jiang and Bansal, 2019). Other models like Qiu et al. (2019) aim at integrating a graph reasoning type of attention where the nodes are recognized by a BERT NER model over the document. Moreover, this model leverages on handcrafted relationships between tokens.

Related to our approach, different papers have investigated the idea of question reformulation to build multi-hop open-domain question answering models. Das et al. (2019) proposes a framework composed of iterative interaction between a document retriever and a reading model. The question reformulation is performed by a *multi-step-reasoner* module trained via reinforcement learning. Similarly, Feldman and El-Yaniv (2019) introduces a multi-hop paragraph retriever. They propose a reformulation component integrated into a retrieving pipeline to iteratively retrieve relevant documents. These works are complementary to ours by focusing mostly on the document retrieving part of the problem while we focus on the answer extraction task, and could be combined together.

This model is designed for a slightly different task called generative question-answering. Moreover, it has been proposed as a way to integrate external knowledge from a database. The major difference with our work is that these architectures propose a reasoning process that sequentially updates the context representation while keeping the representation of the

question unchanged. In addition, we propose to update the representation of the question, reading in the original document all along our pipeline.

4.2 The Latent Question Reformulation Network

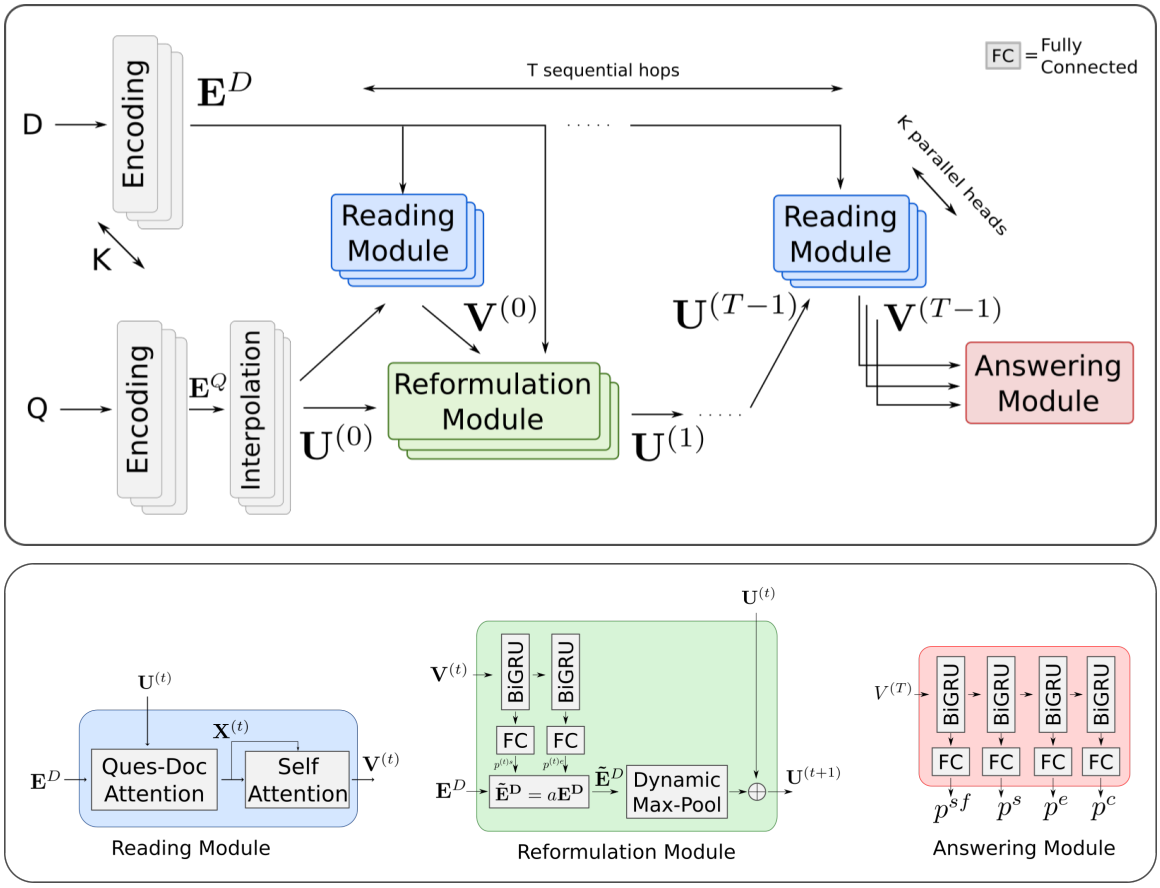

Figure 4.2: Overview of LQR-net with K parallel heads and T sequential reading modules. In this architecture, a latent representation of the question is sequentially updated to perform multi-hop reasoning. K independent reading heads collect pieces of information before feeding them to the answering module. Sections 4.2 present the different building blocks of this end-to-end trainable model.

In this section, we describe the Latent Question Reformulation Network (LQR-net), shown in Figure 4.2. This multi-hop model is designed as an association of four modules: (1) an encoding module, (2) a reading module, (3) a question reformulation module, and (4) an answering module. (1) and (4) are input and output modules, whereas (2) and (3) constitute a hop, and are repeated respectively T and $T-1$ times: the answering module does not require a last reformulation step.

Given a document and a question, the reading module is in charge of computing a question-aware representation of the document. Then, the reformulation module extracts essential elements from this document representation and uses them to update a representation of the question in a latent space. This reformulated question is then passed to the following hop.

The model can have multiple heads, as in the Transformer architecture (Vaswani et al., 2017). In this case, the iterative mechanism is performed several times in parallel in order to compute a set of independent reformulations. The final representations of the document produced by the different heads are eventually aggregated before being fed to the answering module. This module predicts the answer and the *supporting facts* from the document. The following parts of this section describe each module that composes this model.

Note: The model is composed of K independent reading heads that process the document and question in parallel. To not overload the notations of the next parts, we do not subscript all the matrices by the index of the head and focus on the description of one. The aggregation process of the multi-head outputs is explained in Section 4.2.5.

4.2.1 Encoding Module

We adopt a standard representation of each token by using the pre-trained parametric language model BERT (Devlin et al., 2019). Let a document $D = \{p_1, p_2, \ldots, p_{10}\}$ be the set of input paragraphs, of respective lengths $\{n_1, \ldots, n_{10}\}$, associated to a question Q of length L. These paragraphs are independently encoded through the pre-trained BERT model. Each token is represented by its associated BERT hidden state from the last layer

of the model. The tokens representations are then concatenated to produce a global representation of the set of 10 paragraphs of total length $N = \sum_{i=1}^{10} n_i$. The representations are further passed through a Bidirectional Gated Recurrent Unit (BiGRU) (Cho et al., 2014) to produce the final representation of the document $\boldsymbol{E}^D \in \mathbb{R}^{N \times 2h}$ and question $\boldsymbol{E}^Q \in \mathbb{R}^{L \times 2h}$, where h is the hidden state dimension of the BiGRUs.

$$\boldsymbol{E}^Q = \text{BiGRU}(\text{BERT}(Q)), \qquad \boldsymbol{E}^D = \text{BiGRU}([\text{BERT}(p_1); \ldots; \text{BERT}(p_{10})]), \tag{4.1}$$

where $[;]$ is the concatenation operation.

To compute the first representation of the question $\boldsymbol{U}^{(0)}$, we use an interpolation layer to map $\boldsymbol{E}^Q \in \mathbb{R}^{L \times 2h}$ to $\boldsymbol{U}^{(0)} \in \mathbb{R}^{M \times 2h}$ where M is an hyperparameter of the model. Intuitively, $\mathbb{R}^{M \times 2h}$ corresponds to the space allocated to store the representation of the question and its further reformulations. It does not depend on the length of the original question L.

4.2.2 Reading Module

Our model is composed of T hops of *reading* that sequentially extract relevant information from a document regarding the current reformulation of the question. At step t, given a representation of the reformulated question $\boldsymbol{U}^{(t)} \in \mathbb{R}^{M \times 2h}$ and a representation of the document $\boldsymbol{E}^D \in \mathbb{R}^{N \times 2h}$, this module computes a question-aware representation of the document. This module is a combination of two layers: a document-question attention followed by a document self-attention.

Document-Question Attention: We first construct the interaction matrix between the document and the current reformulation of the question $\boldsymbol{S} \in \mathbb{R}^{N \times M}$ as:

$$S_{i,j} = \boldsymbol{w}_1 \boldsymbol{E}^D_{i,:} + \boldsymbol{w}_2 \boldsymbol{U}^{(t)}_{j,:} + \boldsymbol{w}_3 (\boldsymbol{E}^D_{i,:} \odot \boldsymbol{U}^{(t)}_{j,:}), \tag{4.2}$$

where $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$ are trainable vectors of $\mathbb{R}^{2h}$ and $\odot$ the element-wise multiplication.

Then, we compute the document-to-question attention $\boldsymbol{C}^q \in \mathbb{R}^{N \times 2h}$:

$$P_{i,j} = \frac{\exp(S_{i,j})}{\sum_{k=1}^{M} \exp(S_{i,k})}, \quad \boldsymbol{C}^q_{i,:} = \sum_{j=1}^{M} P_{i,j} \boldsymbol{U}^{(t)}_{j,:}. \tag{4.3}$$

And the question-to-document attention $\boldsymbol{q}^c \in \mathbb{R}^{2h}$:

$$m_i = \max_{j \in \{1,\ldots,M\}} S_{i,j}, \quad \boldsymbol{p} = \text{softmax}(\boldsymbol{m}), \quad \boldsymbol{q}^c = \sum_{j=1}^{N} p_j \boldsymbol{E}^D_{j,:}. \tag{4.4}$$

Finally, we compute the question-aware representation of the document $\boldsymbol{X}^{(t)} \in \mathcal{R}^{N \times 8h}$:

$$\boldsymbol{X}^{(t)}_{i,:} = [\boldsymbol{E}^D_{i,:}; \boldsymbol{C}^q_{i,:}; \boldsymbol{E}^D_{i,:} \odot \boldsymbol{C}^q_{i,:}; \boldsymbol{q}^c \odot \boldsymbol{C}^q_{i,:}], \tag{4.5}$$

where [;] concatenation operation. Finally, we use a last BiGRU that reduces the dimension of $\boldsymbol{X}^{(t)}$ to $N \times 2h$. This specific attention mechanism was first introduced in the Bidirectional Attention Flow model of Seo et al. (2017). We hypothesize that such token-level attention will produce a finer-grained representation of the document compared to sentence-level attention used in state-of-the-art Memory Network architectures.

Document Self-Attention: So far, the contextualization between the ten paragraphs has only be done by the BiGRUs of equation 4.1. One limitation of the current representation of the document is that each token has very limited knowledge of the other elements of the context. To deal with long-range dependencies, we apply this same attention mechanism between the question-aware representation of the document, $\boldsymbol{X}^{(t)}$, and itself to produce the reading module output $\boldsymbol{V} \in \mathbb{R}^{N \times 2h}$. This self-contextualization of the document has been found useful in our experiments as presented in the ablation analysis of Section 4.3.3.

4.2.3 Question Reformulation Module

A reformulation module t takes as input the output of the previous attention module $\boldsymbol{V}^{(t)}$, the previous representation of the reformulated question $\boldsymbol{U}^{(t)}$, and an encoding of the document $\boldsymbol{E}^{D}$. It produces an updated reformulation of the question $\boldsymbol{U}^{(t+1)}$.

Reading-based Attention: Given $\boldsymbol{V}^{(t)}$ we compute $\boldsymbol{p}^{(t)s} \in \mathbb{R}^N$ and $\boldsymbol{p}^{(t)e} \in \mathbb{R}^N$ using two BiGRUs followed by a linear layer and a softmax operator. They are computed from:

$$\begin{aligned} \boldsymbol{Y}^{(t)s} &= \text{BiGRU}(\boldsymbol{V}^{(t)}) & \boldsymbol{Y}^{(t)e} &= \text{BiGRU}(\boldsymbol{Y}^{(t)s}) \\ \boldsymbol{p}^{(t)s} &= \text{softmax}(\boldsymbol{w}_s \boldsymbol{Y}^{(t)s}) & \boldsymbol{p}^{(t)e} &= \text{softmax}(\boldsymbol{w}_e \boldsymbol{Y}^{(t)e}), \end{aligned} \tag{4.6}$$

where $\boldsymbol{w}_e$and $\boldsymbol{w}_s$ are trainable vectors of $\mathbb{R}^h$. The two probability vectors $\boldsymbol{p}^{(t)s}$ and $\boldsymbol{p}^{(t)e}$ are not used to predict an answer but to compute a reading-based attention vector $\boldsymbol{a}^{(t)}$ over the document. Intuitively, these probabilities represent the belief of the model at step t of the probability for each word to be the beginning and the end of the answer span. We define the reading-based attention of a token as the probability that the predicted span has started before this token and will end after. It can be computed as follows:

$$a_i^{(t)} = \big(\sum_{k=0}^{i} p_k^{(t)s}\big)\big(\sum_{k=i}^{N} p_k^{(t)e}\big). \tag{4.7}$$

Finally, we use these attention values to re-weight each token of the document representation. We compute $\tilde{E}^{(t)D} \in \mathcal{R}^{N \times 2h}$ with:

$$\tilde{E}_{i,j}^{(t)D} = a_j^{(t)} E_{i,j}^{D}. \tag{4.8}$$

Dynamic Max-Pooling: This layer aims at collecting the relevant elements of $\tilde{\boldsymbol{E}}^{(t)D}$ to add to the current representation of dimension $M \times 2h$. We partition the row of the initial sequence into M approximately equal parts. It produces a grid of $M \times 2h$ in which we apply a max-pooling operator in each individual window. As a result, a matrix of fixed dimension adequately represents the input, preserving the global structure of the document,

and focusing on the important elements of each region. This can be seen as an adaptation of the dynamic pooling layer proposed by Socher et al. (2011).

Formally, let $\tilde{\boldsymbol{E}}^{(t)D}$ be the input matrix representation, we dynamically compute the kernel size, w, of the max-pooling according to the length of the input sequence and the required output shape: $w = \lceil \frac{N}{M} \rceil$, $\lceil \cdot \rceil$ being the ceiling function. Then the output representation of this pooling layer will be $\boldsymbol{O}^{(t)} \in \mathbb{R}^{M \times 2h}$ where

$$O_{i,j}^{(t)} = \max_{k \in \{iw, \ldots, (i+1)w\}} (S_{k,j}). \tag{4.9}$$

Finally, to compute the updated representation of the question $\boldsymbol{U}^{(t+1)} \in \mathbb{R}^{M \times 2h}$, we sum $\boldsymbol{U}^{(t)}$ and $\boldsymbol{O}^{(t)}$.

4.2.4 Answering Module

The answering module is a sequence of four BiGRUs, each of them followed by a fully connected layer. Their respective goal is to supervise (1) the supporting facts p^{sf}, (2) the answer starting and (3) ending probabilities, $\boldsymbol{p}^e$, $\boldsymbol{p}^s$, of each word of the document. (4) The last layer is used as a three-way classifier to predict p^c the probability of the answer be classified as *yes*, *no* or a *span of text*.

$$\begin{array}{lllll} \boldsymbol{Y}^{\text{sf}} = \text{BiGRU}(\boldsymbol{V}^{(t)}) & \boldsymbol{Y}^s = \text{BiGRU}(\boldsymbol{Y}^{\text{sf}}) & \boldsymbol{Y}^e = \text{BiGRU}(\boldsymbol{Y}^s) & \boldsymbol{Y}^c = \text{BiGRU}(\boldsymbol{Y}^e) \\ & \boldsymbol{p}^s = \text{softmax}(\boldsymbol{w}_s \boldsymbol{Y}^s) & \boldsymbol{p}^e = \text{softmax}(\boldsymbol{w}_e \boldsymbol{Y}^e) & \boldsymbol{p}^c = \text{softmax}(\boldsymbol{w}_c \boldsymbol{Y}^c) \end{array} \tag{4.10}$$

where $\boldsymbol{w}_s \in \mathbb{R}^h, \boldsymbol{w}_e \in \mathbb{R}^h, \boldsymbol{W}_c \in \mathbb{R}^{h \times 3}$ are trainable parameters.

To predict the supporting facts, we construct a sentence based representation of the document. Each sentence is represented by the concatenation of its starting and ending supporting fact tokens from $\boldsymbol{Y}^{\text{sf}}$. We compute $p_{i,j}^{\text{sf}}$ the probability of sentence j of example i of being a supporting fact with a linear layer followed by a sigmoid function.

4.2.5 Multi-head Version

We define a multi-head version of the model. In this configuration, we use a set of independent parallel heads. All heads are composed of the same number of reading and reformulation modules. Each head produces a representation $\boldsymbol{V}_k^{(T)}$ of the document. We finally sum these K matrices to compute the input of the answering block.

4.2.6 Training

We jointly optimize the model on the three subtasks (supporting facts, span position, classifier *yes/no/span*) by minimising a linear combination of the supporting facts loss $\mathcal{L}_{\text{sf}}$, the span loss $\mathcal{L}_{\text{span}}$ and the class loss $\mathcal{L}_{\text{class}}$. Let N_d be the number of examples in the training dataset. $\mathcal{L}_{\text{sf}}(\theta)$ is defined by:

$$\mathcal{L}_{\text{sf}}(\theta) = \frac{1}{N_d} \sum_{i}^{N_d} \frac{1}{\text{nbs}_i} \sum_{j}^{\text{nbs}_i} (p_{i,j}^{\text{sf}} - y_{i,j}^{(1)})^2, \tag{4.11}$$

where nbs_i corresponds to the number of sentences in the document i. $y_{i,j}^{(1)}$ being 1 if the sentence j of the document i is a supporting fact otherwise 0.

Selecting the answer in multi-hop reading datasets is a weakly supervised task. Indeed, similarly to the observations of Min et al. (2019a) for open-domain question-answering and discrete reasoning tasks, it is frequent for a given answer of HOTPOTQA to appear multiple times in its associated document. In our case, we assume that all the mentions of the answer in the supporting facts are related to the question. We tag as a valid solution, the start and end positions of all occurrences of the answer in the given supporting facts.

$\mathcal{L}_{\text{span}}(\theta)$ is defined by:

$$\mathcal{L}_{\text{span}}(\theta) = \frac{1}{N_d} \sum_{i}^{N_d} \frac{1}{2} D_{\text{KL}}(p_i^s \| y_i^{(2)}) + D_{\text{KL}}(p_i^e \| y_i^{(3)}) \tag{4.12}$$

where $y_i^{(2)} \in \mathbb{R}^N, y_i^{(3)} \in \mathbb{R}^N$ are vectors containing the value $1/n_i$ at the start, end positions

of all the occurrences of the answer, 0 otherwise; n_i being the number of occurrences of the answer in the context.

$\mathcal{L}_{\text{class}}(\theta)$ is defined by:

$$\mathcal{L}_{\text{class}}(\theta) = -\frac{1}{N_d}\sum_{i}^{N_d}\log(p^c_{i,y_i^{(4)}}), \tag{4.13}$$

where $y_i^{(4)}$ corresponds to the index of the label of the question type {*yes*, *no*, *span*}. We finally define the training loss as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{class}}(\theta) + \alpha\mathcal{L}_{\text{span}}(\theta) + \beta\mathcal{L}_{\text{sp}}(\theta), \tag{4.14}$$

where α and β are hyperparameters tuned by cross-validation.

4.3 Experiments

4.3.1 Data Augmentation

In the original HOTPOTQA dataset, the two *gold* paragraphs required to answer a given question come with eight distractor paragraphs. These eight distractor paragraphs, collected from Wikipedia, are selected among the results of a bigram tf-idf retriever (Chen et al., 2017) using the question as the query. As an augmentation strategy, we created additional "easier" examples by combining the two *gold* paragraphs with eight other paragraphs randomly selected in the dataset. For each example of the original training set, we generate an additional "easier" example. These examples are shuffled in the dataset.

4.3.2 Implementation Details

Our model is composed of 3 parallel heads ($K = 3$) each of them composed of two reading modules and one reformulation module ($T = 2$). We set the hidden dimension of all the GRUs to $d = 80$. We use $M = 100$ to allocate a space of $\mathbb{R}^{100\times160}$ to store the question

and its reformulations. We use pre-trained *BERT-base-cased* model (Devlin et al., 2019) and adapt the implementation of *Hugging Face*[1] to compute embedding representations of documents and questions. We optimize the network using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of $1e^{-4}$. We set α to 1 and β to 10. All these parameters have been defined through cross-validation.

4.3.3 Results and Ablation Analysis

	Model	Answer		Sup Fact		Joint	
		EM	F_1	EM	F_1	EM	F_1
Contemporary Works	Longformer (Beltagy et al., 2020)	68.00	81.25	63.09	88.34	45.91	73.16
	C2F Reader (Shao et al., 2020)	67.98	81.24	60.81	87.63	44.67	72.73
	HGN (Fang et al., 2020)	66.07	79.36	60.33	87.33	43.57	71.03
	TAP2 (Bhargav et al., 2020)	64.99	78.59	55.47	85.57	39.77	69.12
	SAE (Tu et al., 2020)	60.36	73.58	56.93	84.63	38.81	64.96
	LQR-net (ours) (Grail et al., 2020)	60.20	73.78	56.21	84.09	36.56	63.68
	DFGN (Qiu et al., 2019)	56.31	69.69	51.50	81.62	33.62	59.82
	QFE (Nishida et al., 2019)	53.86	68.06	57.75	84.49	34.63	59.61
	Baseline Model (Yang et al., 2018)	45.60	59.02	20.32	64.49	10.83	40.16
	Unsupervised Decomposition (Perez et al., 2020)	66.33	79.34	-	-	-	-
	ChainEx (Chen et al., 2019)	61.20	74.11	-	-	-	-
	DecompRC (Min et al., 2019b)	55.20	69.63	-	-	-	-
	Self-Assembling NMN (Jiang and Bansal, 2019)	49.58	62.71	-	-	-	-

Table 4.1: Performance comparison on the private test set of HOTPOTQA in the distractor setting. We compare our model, in term of Exact Match and F_1 scores, against the *published* (code or paper) single models. Our submission is tagged as *LQR-net 2 + BERT-Base (single model)* on the official leaderboard (https://hotpotqa.github.io/). First part of the table shows models that are contemporary to ours.

Table 4.1 presents the performance of our LQR-net on the distractor setting of the HOTPOTQA dataset. We compare our model against the *published* approaches evaluated on the HOTPOTQA dataset. We can see from this table that our model achieves strong performance on the answer prediction task. It outperforms the best model, non contemporary to

[1] https://github.com/huggingface/pytorch-transformers

this work, by 3.9 points of EM and 4.1 points of F_1 score. Our model also achieves competitive performance for the evidence extraction task. The LQR-net achieves state-of-the-art performance on the joint task improving the best *published*, non contemporary approaches by 2.9 points on EM and 3.9 points of F_1.

Model	Answer		Sup Fact		Joint	
	EM	F_1	EM	F_1	EM	F_1
LQR	**60.0**	**74.1**	**55.8**	83.9	**36.5**	**64.0**
- Data aug	59.3	73.4	52.8	**84.2**	34.4	63.6
CE Loss	59.6	73.6	52.7	83.5	34.4	63.2
K = 1	59.2	73.2	48.9	83.8	31.6	63.0
- Self-Att	53.4	66.8	48.9	79.2	30.1	55.7
T = 1	53.4	67.2	48.3	78.2	28.8	55.1
M = 1	51.8	65.2	42.1	72.1	25.8	50.7

Table 4.2: Comparison of different architectures and model choices against the best configuration on the development set of HotpotQA.

To evaluate the impact of the different components of our model, we perform an ablation analysis. Table 4.2 presents the results of this analysis.

Impact of sequential and parallel reading: We study the contributions of the sequentiality in the model and of the multiple parallel heads. We compare our model to a similar architecture without the sequential reformulation ($T = 1$). We find that this sequential association of reading modules and reformulation modules is a critical component. F_1 score decreases by 6.9 points for the answer prediction task and 5.7 points for the evidence extraction task when the model does not have the capability to reformulate the question.

The impact of the parallel heads is more limited than the sequentiality but still remains significant. Indeed, the configuration that uses only a single head ($K = 1$) stands 1 F_1 points below the best model on the joint metric.

These results lead us to conclude that sequential and parallel reading are required to deal with the two types of reasoning presented in at the beginning of the chapter.

Weak supervision of the answer: In this work, we propose to label as positive all occurrences of the answer in the supporting facts. We compare this configuration to the standard approach, where only the first occurrence of the answer is labeled as positive and the others as negative. In this last configuration, the span loss corresponds to a cross-entropy loss (CE loss) between the predicted start and end probabilities and the target positions. This decreases the joint F_1 score by 0.8 points.

Impact of the self-attention layer: We study the impact of the self-attention layer in the reading module. We found that this self-attention layer is an essential component in the reading process. Indeed, when we omit this layer, the F_1 score decreases by 8.3 points on the joint metric. This outlines the necessity to be able to propagate long-range information between the different paragraphs and not only in the local neighborhood of a token. Compared to previously proposed approaches, this layer does not rely on any handcrafted relationship across words.

Question as a single vector: Finally, we study the case where the question representation is reduced to a vector of $\mathbb{R}^{2h}$ ($M = 1$). This configuration achieves the worst results of our analysis, dropping the joint F_1 score by 13.3 points and highlights the importance of preserving a representation of the question as a matrix to maintain its meaning.

4.3.4 Open-Domain Experiments

In this part, we describe how we integrated our model into an entire reading pipeline for open-domain question answering. In this setting, no supporting documents are associated to each question, and it is required to retrieve relevant context from large text corpora such as Wikipedia. We adopt a two-stage process, similar to Chen et al. (2017); Clark and Gardner (2018), to answer multi-hop complex questions based on the 5 million documents of Wikipedia. First, we use a paragraph retriever to select a limited amount of relevant paragraphs from a Wikipedia dump, regarding a natural language question. Second, we fed our LQR model with the retrieved paragraphs to extract the predicted answer. We evaluate this approach on the open-domain configuration of the HotpotQA dataset called ***fullwiki***.

	Model	Answer		Sup Fact		Joint	
		EM	F_1	EM	F_1	EM	F_1
Contemporary Works	HopRetriever + Sp-search (Li et al., 2020)	62.1	75.2	52.5	78.9	37.8	64.5
	IRRR+† (Qi et al., 2020)	66.33	79.10	56.92	83.24	42.75	69.60
	Recursive Dense Retriever (Xiong et al., 2020)	62.28	75.29	57.46	80.86	41.78	66.55
	DDRQA (Zhang et al., 2020b)	62.9	75.9	51.3	79.1	-	-
	Robustly Fine-tuned GRR (Asai et al., 2020)	65.8	52.7	75.0	47.9	-	-
	Transformer-XH-final (Zhao et al., 2020)	54.0	66.2	41.7	72.1	27.7	52.9
	SemanticRetrievalMRS (Nie et al., 2019)	**46.50**	**58.80**	**39.90**	**71.5**	**26.6**	**49.2**
	LQR-net (ours) (Grail et al., 2021)	43.00	54.0	30.10	58.90	18.90	39.20
	GOLDEN Retriever† (Qi et al., 2019)	37.92	48.58	30.69	64.4	18.04	39.13
	CogQA (Ding et al., 2019)	37.60	49.40	23.10	58.5	12.2	35.3
	MUPPET (Feldman and El-Yaniv, 2019)	31.07	40.42	17.00	47.71	11.76	27.62
	QFE† (Nishida et al., 2019)	28.70	38.10	14.20	44.40	8.69	23.1
	Baseline Model (Yang et al., 2018)	24.68	34.36	5.28	40.98	2.54	17.73
	DecompRC (Min et al., 2019b)	-	43.26	-	-	-	-

Table 4.3: Performance comparison on the development set of HOTPOTQA in the *fullwiki* setting. We compare our model in terms of Exact Match and F_1 scores against the *published* (code or paper) models. † indicates that the paper does not report the results on the development set of the dataset; we display their results on the test set. First part of the table shows state-of-the art methods that are contemporary to ours.

We use a standard TF-IDF based paragraph retriever to retrieve the paragraphs the most related to the question. In addition to these paragraphs, we consider as relevant their *neighbors* in the Wikipedia graph, i.e. the documents linked to them by hyperlinks. In our experiments, we considered as relevant, the top 10 paragraphs and their associated *neighbors*.

Table 4.3 shows the results of our approach compared to other published models. Although we are using a very simple retriever, only based on TF-IDF, we report competitive results on the open-domain question answering task of HotpotQA. The only non contemporary published approach (Nie et al., 2019) that outperforms us being a combination of sentence/paragraph retrieval based on BERT encodings.

4.3.5 Qualitative Analysis

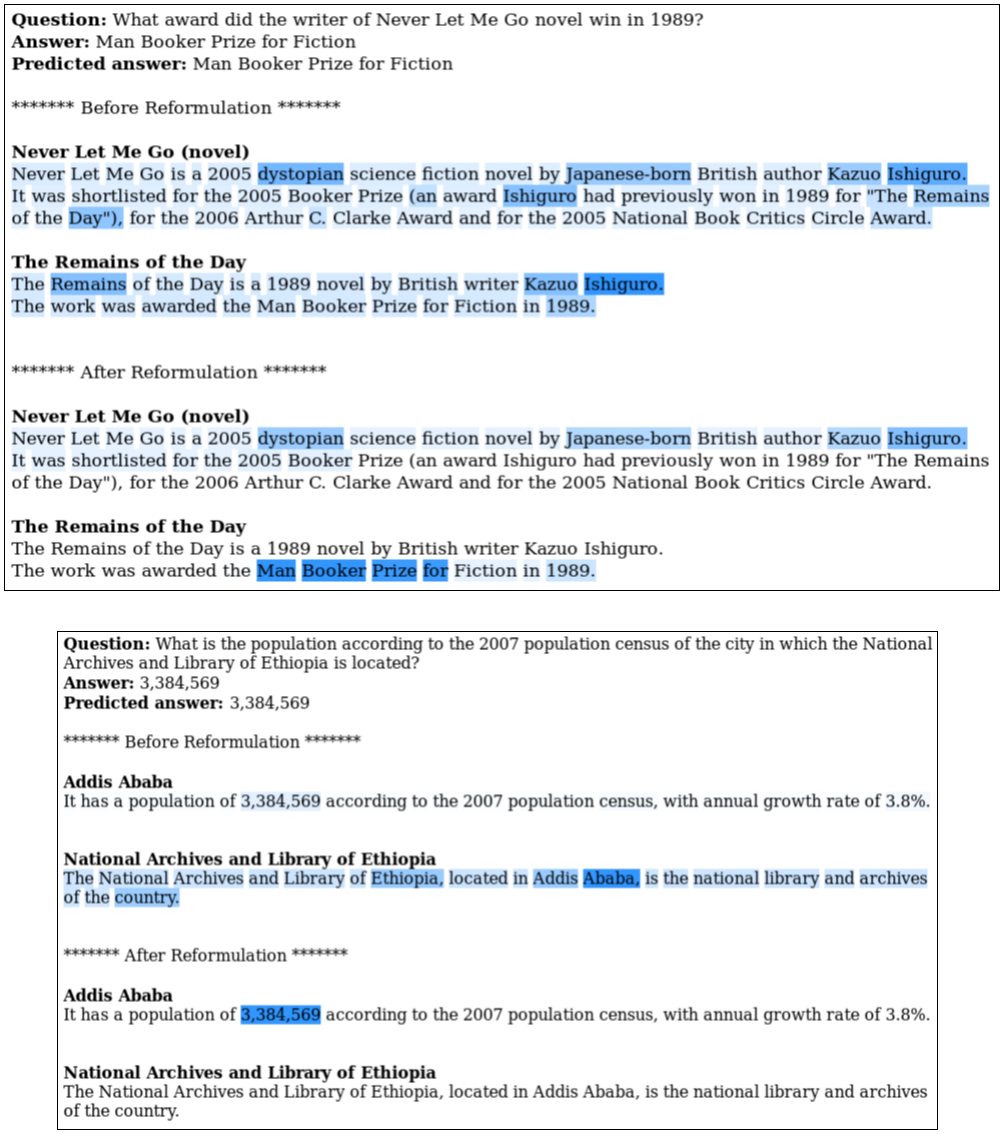

Figure 4.3: Distribution of the probabilities for each word to be part of the predicted span, before the first reformulation module and in the answering module. We display the reading-based attention computed in Equation 4.7 and the reading-based attention computed from $\boldsymbol{p}^s$ and $\boldsymbol{p}^e$ from Equation 4.10. In these examples, we show only the supporting facts.

Question Reformulation and Reasoning Chains: Because our model reformulates the question in a latent space, we cannot directly visualize the text of the reformulated question. However, one way to assess the effectiveness of this reformulation is to analyze the evolution of $\boldsymbol{p}^s$ and $\boldsymbol{p}^e$ across the two hops of the model. We present in Figure 4.3 an analysis of the evolution of these probabilities on two *bridge* samples of the development dataset. We display the reading-based attention, that corresponds to the probabilities for each word to be part of the predicted span, computed from $\boldsymbol{p}^s$ and $\boldsymbol{p}^e$ in Equation 4.7. These examples show this attention before the first reformulation of the question and in the answering module.

From these observations, we can see that the model tends to follow a natural reasoning path to answer *bridge* questions. Indeed, before the first reformulation module, the attentions tend to focus on the first step of reasoning. For the question *"What award did the writer of Never Let Me Go novel win in 1989?"*, the model tends to focus on the name of the writer at the first step, before jumping the award description in the second step. Similarly, for the question *"What is the population according to the 2007 population census of the city in which the National Archives and Library of Ethiopia is located?"* we can see the model focusing on Addis Ababa at the first step, i.e the name of the city where the National Archives and Library of Ethiopia are located and then jumping to the population of this city in the next hop.

Limitations: We manually examine one hundred errors produced by our multi-step reading architecture on the development set of HOTPOTQA . We identify three recurrent cases of model failure: (1) the model stops at the first hop of required reasoning, (2) the model fails at comparing two properties, and (3) the answer does not match all the requirements of the question. We illustrate these three recurrent types of error with examples from the dataset in Figure 4.4.

The answer does not match all the requirements of the question:

Question: Who is younger, Wayne Coyne or Toshiko Koshijima?
Answer: Toshiko Koshijima
Predicted answer: wayne michael coyne)

Wayne Coyne
✱ Wayne Michael Coyne(born January 13, 1961) is an American musician.
He is the lead singer, occasional backing vocalist, guitarist, keyboardist, theremin player and songwriter for the band the Flaming Lips.

Toshiko Koshijima
✱ Toshiko Koshijima こしじま としこ , Koshijima Toshiko ,born March 3, 1980 in Kanazawa, Ishikawa) is a Japanese singer.
Along with composer, record producer and DJ Yasutaka Nakata, she is a lead vocalist of the electronica band Capsule, which they formed in 1997 when both were 17.
Their formal debut came in 2001 with the release of the single "Sakura".
Two more singles and their debut album, "High Collar Girl", followed the same year.

The answer does not match all the requirements of the question:

Question: Approximately how many locations is BJ's Wholesale Club operating in, as of early 2008?
Answer: 650 locations
Predicted answer: 500

US Vision
U.S. Vision, a wholly owned subsidiary of Refac Optical Group, is an international optometric dispensary chain.
The vast majority of these locations are leased spaces in large department stores, such as J.C. Penney, Boscov's, and The Bay.
As of May 8, 2007, 500 locations in 47 states and Canada are in operation, consisting of licensed departments and freestanding stores.
✱ In early 2008, due to an acquisition of BJ's Optical Centers located in many BJ's Wholesale Clubs, that number has grown to approximately 650 locations.
U.S. Vision deals mainly in prescription eyewear, contact lenses, and optometry offices.

BJ's Wholesale Club
✱ BJ's Wholesale Club Inc., commonly referred to simply as BJ's, is an American membership-only warehouse club chain operating on the United States East Coast, as well as in the state of Ohio.

The answer does not match all the requirements of the question:

Question: What is one of the stars of The Newcomers known for?
Answer: superhero roles as the Marvel Comics
Predicted answer: chris evans

Chris Evans (actor)
✱ Christopher Robert Evans (born June 13, 1981) is an American actor and filmmaker.
Evans is known for his superhero roles as the Marvel Comics characters Steve Rogers / Captain America in the Marvel Cinematic Universe and Johnny Storm / Human Torch in "Fantastic Four" and .

The Newcomers (film)
✱ The Newcomers is a 2000 American family drama film directed by James Allen Bradley and starring Christopher McCoy, Kate Bosworth, Paul Dano and Chris Evans.
Christopher McCoy plays Sam Docherty, a boy who moves to Vermont with his family, hoping to make a fresh start away from the city.
It was filmed in Vermont, and released by Artist View Entertainment and MTI Home Video.

Figure 4.4: Examples from the HOTPOTQA development set that illustrate the categories of errors presented in Section 4.3.5. For each example, we show only the text of the two gold paragraphs. Supporting facts are identified with $*$.

During this analysis of errors, we found that in only 3% of the cases, the answer is selected among one of the distractor paragraphs instead of a *gold* one. Our architecture successfully detects the relevant paragraphs regarding a question even among similar documents coming from a tf-idf retriever. Moreover, there are no errors where the model produces a binary *yes/no* answer instead of extracting a text span and vice versa. Identifying the type of question is not challenging for the model. This might be explained by the question's "patterns" that are generally different between binary *yes/no* and extractive questions.

Summary of our Contributions to Question-Answering

In this first part of this thesis, we describe our contributions related to adversarial reading and multi-hop question-answering. Chapter 3 describes our proposed adversarial learning protocol based on a couple of competitive reading models. This adversarial protocol iteratively creates corrupted examples, derived from the original ones, that increase the task's difficulty. We demonstrate the effectiveness of this adversarial protocol by experimenting it with several models and datasets. We compare it to a *standard* training procedure and a training based on uniform corruption of data. Our approach brings performance improvement compared to both baseline protocols. Finally, we propose several visualizations that allow interpreting how the reader produces an answer, and which parts of the document are crucial for it to make its decision.

In the second chapter of this part, we propose a competitive and interpretable model to deal with multi-hop question-answering. We propose the Latent Question Reformulation Network, a reading architecture that sequentially reformulates and answer a given question. We evaluate our model on HOTPOTQA , a multi-hop question-answering dataset based on Wikipedia data, and show its effectiveness in extracting meaningful answers from a set of documents. Besides being competitive, our approach produces an interpretable decomposition of the reasoning path of the model illustrated on several examples.

Part II

Learning Long Document Representations

Chapter 5

A scalable Transformer architecture for summarizing long documents

As presented in Section 2.1.2, language model pre-training has become a key component to improve performances on a majority of Natural Language Processing tasks (Wang et al., 2019a). Most of the recent competitive architectures (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019b; Radford et al., 2018) are based on the efficient transformer layer introduced in Vaswani et al. (2017). BERT (Devlin et al., 2019) is one of these architectures that has been widely adopted for comprehension and generation tasks. It is a multi-layer transformer network, pre-trained with different self-supervised objectives. Numerous variations of transformer architectures have been proposed to improve this approach (Lan et al., 2020; Liu et al., 2019b; Radford et al., 2018). However, this type of process is only evaluated on tasks composed of relatively short input text, GLUE (Wang et al., 2019a), SQUAD (Rajpurkar et al., 2016), SWAG (Zellers et al., 2018). Indeed, for the tasks that require reasoning with longer documents, this approach exhibits several limitations. The transformer self-attention memory quadratically increases with the number of input tokens, making it technically impossible to compute on document-scale sequences. In addition, they usually require to define a fixed maximum input length, typically of 512 tokens, at the pre-training stage.

One solution is to pre-train the entire model on longer sequences. However, this will still require a massive computation power and will only push the length limitation further. Other alternatives have been proposed to extend multi-layer transformers architectures to longer sequences without modifying this maximum length limitation. The first one is to limit the input sequence to its first tokens by removing the text beyond the length limit. Obviously, it cannot be a reasonable solution to treat long documents that are consistently longer than this limit. The second alternative is to apply the model on a window that slides all over the document. It has been used in Wolf et al. (2019) to deal with SQUAD documents that are longer than the 512 token limitation and in Joshi et al. (2019) for a co-reference resolution task on long documents. This approach can only work if the tokens need to be contextualized only in their surroundings because there is no interaction between the different windows. It seems to be a solution for co-reference resolution (Joshi et al., 2019) as they usually can be solved with a reasonably sized window. Another approach adopted to deal with long documents or multi-document is to select a sub-sample of the input that is small enough for the transformer model. Most of the state-of-the-art pipelines on the multi-hop question answering dataset HotpotQA (Yang et al., 2018) use a first model to retrieve the relevant pieces of text before feeding them to a transformer-based architecture (Fang et al., 2020; Tu et al., 2020).

We argue that these solutions are not feasible to deal with tasks that require a global understanding of long documents. An example is extractive summarization, where the decision for each sentence should be based on the information of the complete document. To address these challenges, we propose a simple adaptation of the multi-layer transformer architecture that can scale to long documents and benefit from pre-trained parameters with a relatively small length limitation. The general idea is to independently apply a transformer network on small blocks of a text, instead of a long sequence, and to share information among the blocks between two successive layers. To the best of our knowledge, this is the first attempt to introduce hierarchical components directly between the layers of a pre-trained model and not only on top of it (Fang et al., 2020; Zhang et al., 2019; Tu et al., 2020). Between each of the transformer layers, we use a Bidirectional Gated Recurrent Unit

(BiGRU) network (Cho et al., 2014) to spread global information across the blocks. Adding these propagation layers between the transformer layers preserves the original structure of the pre-trained model and makes it possible to transfer parameter weights from a large pre-trained language model with only few additional parameters to propagate information between blocks.

The contributions of this chapter can be summarized as follows: (i) we propose a novel architecture dedicated to long documents which interweaves recurrent hierarchical modules with transformer layers and which exploits pre-trained language models like BERT, and (ii) we demonstrate that this architecture constructs informative representations in the context of extractive summarization and long document matching. This chapter extends our work published in Grail et al. (2021) during the thesis.

5.1 Related Work

Hierarchical neural architectures have been competitive on a collection of NLP tasks that require to reason over long or multiple documents such as aspect-based sentiment analysis (Paulus et al., 2018), document summarization (Li et al., 2015; Cheng and Lapata, 2016b), document segmentation (Koshorek et al., 2018) and text classification (Yang et al., 2016). The hierarchical structure enables the model to learn local contextualized token representations in its lower hierarchy level, while higher-level representations can capture long-distance dependencies within the document. Liu and Lapata (2019a) have proposed a hierarchical modification of the transformer layer-based attention modules to model relations between documents for abstractive summarization but do not investigate parameter transfer form pre-trained language models. Chang et al. (2019) and Zhang et al. (2019) suggested pre-training processes for hierarchical models, without however testing their approaches on long document summarization nor releasing their pre-trained models. We have not included these models in our comparison for this reason. Transformer-XH (Zhao et al., 2020) introduced an eXtra Hop attention to model dependencies between different transformer windows but requires a graph of related documents.

Long-Document Transformers: There has been a large variety of approaches recently proposed to extend the Transformer model to long documents. The simple and straightforward sliding windows method has been proposed by Wang et al. (2019b). In this paper, articles are split into passages with a length of 100 words, and a Transformer model is applied within each passage. There is no connection between the windows. This approach has several drawbacks. The main one being that it does not handle long-term dependencies because information does not flow beyond a single window. To overcome this issue, this approach is often combined with a strategy of token selection as detailed after.

One of the popular categories of efficient Transformer is related to auto-regressive approaches. Transformer-XL (Dai et al., 2019) and XLNet (Yang et al., 2019) proposed a recurrence between successive transformer windows which run from left-to-right through the document. With these approaches, the model is able to use information from the past to compute efficient representations of future tokens. Rae et al. (2020) introduced an improvement over this work with the Compressive Transformer, which has access to compressed memories that represent tokens of the past. Auto-regressive models work well from left-to-right language models but suffer in tasks that require bidirectional context.

Other approaches have designed the self-attention as a sparse layer. This includes works such as sparse Transformer (Child et al., 2019), Longformer (Beltagy et al., 2020), BIGBIRD, (Zaheer et al., 2020), the Blockwise Transformer (Qiu et al., 2020). Reformer (Kitaev et al., 2020) also tackles the problem of expensive self-attention in the context of long document. It proposes a sparse attention computed only between *similar* tokens, based on locality-sensitive hashing.

A recent direction has emerged to build efficient Transformer models and is related to low-rank, and kernels approaches. This line of work includes Linear Transformer (Katharopoulos et al., 2020), Lineformer (Wang et al., 2020), Performer (Choromanski et al., 2021). These papers propose several approximations of the self-attention that linearly scale with the number of input tokens.

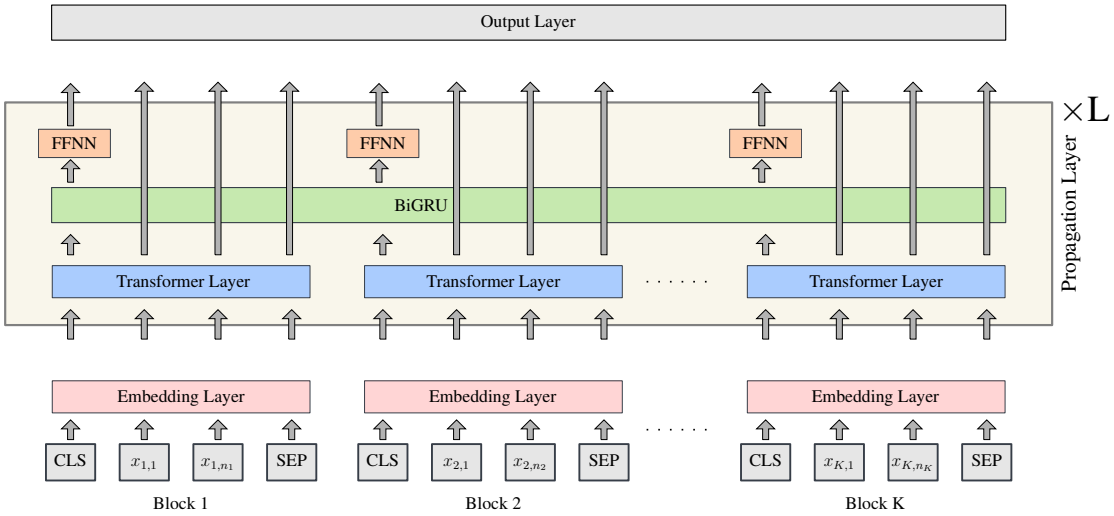

Figure 5.1: Our proposed modification of a multi-layer transformer architecture. The input sequence is composed of K blocks of tokens. Each transformer layer is applied within the blocks, and a bidirectional GRU network propagates information in the whole document by updating the [CLS] representation of each block.

Our proposed approach is related to sparse methods with an attention computed only between specific tokens. It is also related to auto-regressive methods with access to information from other windows. However, contrary to the presented auto-regressive approaches, our model can access information that flows from left-to-right and also from right-to-left of the document.

5.2 Globalizing BERT-based Architecture

In this part, we briefly recall the the BERT architecture (Devlin et al., 2019) which is an essential building block of our proposition. Then we describe our modifications of this architecture that allow the model to read longer documents.

BERT (Devlin et al., 2019) is a multi-layer transformer encoder pre-trained on large text corpora. Two BERT architectures have been proposed in Devlin et al. (2019): $\text{BERT}_{\text{BASE}}$ composed of 12 stacked transformer layers with hidden dimension of 768 ($L = 12, h = 768$) and $\text{BERT}_{\text{LARGE}}$ composed of 24 layers of hidden dimension 1024 ($L = 24, h =$

1024). For both architectures, the input length is limited to 512 WordPiece tokens and the pre-training includes two self-supervised tasks, namely masked language modeling and next sentence prediction. For masked language modeling, 15% of all the WordPiece tokens of the input sequence are masked or corrupted, and the model is used to predict the original token with a cross-entropy loss. For next sentence prediction, the model is trained as a classifier to predict if two sentences are contiguous or not. The pre-training procedure uses the BooksCorpus (Zhu et al., 2015) and documents from English Wikipedia. It requires 4 days of optimization on 16 TPU chips for $\text{BERT}_{\text{BASE}}$ and 64 TPU chips for $\text{BERT}_{\text{LARGE}}$.

5.2.1 Stacked Propagation Layers

We propose a hierarchical structure that uses pre-trained transformers to encode local text blocks that will be used to compute document level representations. The novel contribution of this work, depicted in Figure 5.1, is to incorporate recurrent hierarchical modules between the different transformer layers and not only on top of the model, as proposed in several recent works (Fang et al., 2020; Zhang et al., 2019; Tu et al., 2020). Because we construct and propagate document level information between the layers, global and local information are fused at every level of the architecture. The text blocks can be sentences, paragraphs, or sections. We experiment using sentences as blocks because it generally does not exceed the maximum length allowed by pre-trained models and because BERT has demonstrated to be well adapted to represent such sequences.

We start by splitting the original sequence into multiple blocks. Let D be a document composed of K blocks, $D = \{B_1; B_2; \cdots; B_K\}$ where a block B_k, $1 \leq k \leq K$, is composed of n_k tokens. To follow the convention of BERT, special tokens [CLS] and [SEP] are respectively added at the beginning and end of each block of the document, so that: $B_k = \{[\text{CLS}]; x_{k,1}; x_{k,2}; \cdots x_{k,n_k}; [\text{SEP}]\}$ where $x_{k,i}$ is the index of the WordPiece token i of block k. In the remainder, the index 0 (resp. $n_k + 1$) will be used to refer to the representation of the [CLS] (resp. [SEP]) token in each block.

Embedding Layer Because our goal is to reuse the available pre-trained BERT parameters, token representations are kept the same as in the original BERT and are composed of a token embedding, a segment embedding, and a positional encoding that represents the position of the token in its block. We will denote by E_k ($E_k \in \mathbb{R}^{(n_k+2)\times h}$, $1 \leq k \leq K$) the embedding representation of block k.

Propagation Layers Our model is composed of L stacked identical hierarchical layers, called *propagation layers*, that comprise a transformer layer, a BiGRU to propagate information across blocks and, finally, a feed-forward network. For any layer ℓ, $1 \leq \ell \leq L$, let $U_k^\ell \in \mathbb{R}^{(n_k+2)\times h}$ be the representation of block k after the $(\ell-1)^{th}$ layer, the representation for the first layer being initialized with the output of the embedding layer: $U_k^1 = E_k, \quad \forall k \in \{1, \cdots, K\}$. We first apply the pre-trained transformer function T^ℓ individually on each block of the document to compute local, token-aware representations $V_k^\ell \in \mathbb{R}^{(n_k+2)\times h}$:

$$V_k^\ell = T^\ell(U_k^\ell), \quad \forall k \in \{1, \cdots, K\}.$$

The next step is to propagate information across all the blocks of the document in order to compute a global block-aware representation for the document at layer ℓ, denoted by $W^\ell \in \mathbb{R}^{K\times h}$, $1 \leq k \leq K$. To do so, we use a BiGRU network, fed with the representation vectors of the different blocks, and apply a feed-forward neural network to preserve the hidden dimension of the transformer. Each block k is represented by its [CLS] vector, *i.e.*, the vector (represented by $V_{k,0}^\ell \in \mathbb{R}^h$) at the first position in the local representation of the block. These representations are then concatenated to form the input to the BiGRU. The global, block-aware representation is then computed by applying the feed-forward neural network (FFNN) to all K outputs of the BiGRU:

$$W_k^\ell = \text{FFNN}(\text{BiGRU}_k([V_{1,0}^\ell; \cdots; V_{K,0}^\ell])),$$

where BiGRU_k denotes the k^{th} output of the BiGRU and $[;]$ is the concatenation operation.

At this stage, we have computed, for a given document, local block representations V_k^ℓ $(1 \leq k \leq K)$ and a global representation W^ℓ. We combine them to build the output

representation of the layer:

$$U_k^{\ell+1} = [W_k^{\ell}; V_{k,1}^{\ell}; \cdots ; V_{k,n_k+1}^{\ell}], \quad 1 \leq k \leq K.$$

As one can note, $U_k^{\ell+1} \in \mathbb{R}^{(n_k+2)\times h}$ is a representation of block k in which the [CLS] vector representation has been enriched with document level information propagated from other blocks. $U_k^{\ell+1}$ is then used as input for the next propagation layer.

5.2.2 Output Layer

In this work, we validate our approach on tow diferent tasks: extractive summarization and long document matching. Section 5.3 describes these tasks in details.

Extractive Summarization: We consider the extractive summarization as a binary classification problem where each block has to be labeled as selected or not. We use a feed-forward neural network followed by a Softmax function on the top of the block level representations after the last layer L to compute $Y \in \mathbb{R}^{K\times 2}$.

$$Y_k = \text{Softmax}(\text{FFNN}(W_k^{L+1})).$$

Long-to-long Document Matching: Following Yang et al. (2020) we use a siamese architecture to compute the similarity between two documents. In this configuration, the output layer of G-BERT needs to produce a document representation $Y \in \mathbb{R}^o$, o being the output dimension. To do so, we concatenate the representation of the first and last block of the document and feed them to a feed-forward neural network.

$$Y = \text{FFNN}([W_0^{L+1}; W_K^{L+1}])$$

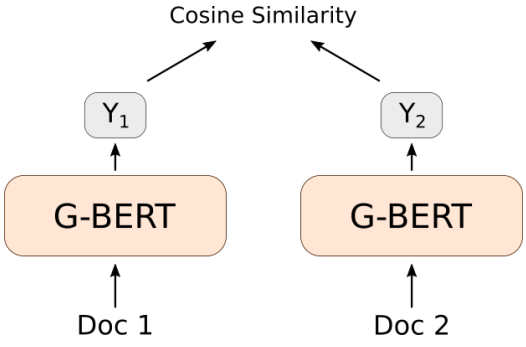

Figure 5.2: Siamese configuration of G-BERT for long document matching.

Datasets	avg. doc length		avg. summary length	
	sentences	words	sentences	words
arXiv	204	5038	5.6	165
PubMed	88	3235	6.8	205
MultiNews	82	2094	9.9	256
CNN/DM	32	757	4.1	57

Table 5.1: Statistics on arXiv, PubMed, MultiNews and CNN/DailyMail validation datasets in terms of documents and summary lengths.

Using a recurrent architecture to propagate information between blocks has two interesting properties. First, it allows our model to scale to long sequences of blocks without using an attention mechanism that would not scale. Second, it does not require to implement any positional encoding on block representations.

5.3 Extractive Summarization Experiments

5.3.1 Dataset Description

We first evaluate our approach, which we refer to as G-BERT (for 'Global BERT-based architecture'), in the context of extractive summarization.

The goal of extractive summarization is to identify and extract from a document the pieces of text that are the most important (Kupiec et al., 1995a). We view this task as a sentence-level classification problem where each sentence has to be labeled according to its belonging to the summary or not. To validate the effectiveness of our approach, we propose to test it on four summarization datasets, namely ArXiv, PubMed, Multi-News and CNN/DailyMail:

- The **ArXiv** and **Pubmed** datasets have been introduced in Cohan et al. (2018). They contain long scientific documents from `arXiv.org` and `PubMed.com` and use their abstracts as the ground-truth summaries. We use the original splits that respectively contain $203,037/6,436/6,440$ samples in the training, validation, and test sets

		PubMed				arXiv			
	Summarizer	RG-1	RG-2	RG-3	RG-L	RG-1	RG-2	RG-3	RG-L
	Oracle	58.15	34.16	24.11	52.99	57.78	30.43	18.41	51.24
	Lead	37.77	13.35	7.64	34.31	35.54	9.50	3.33	31.19
Abstractive or Mix	Attn-Seq2Seq (Nallapati et al., 2016a)	31.55	8.52	7.05	27.38	29.30	6.00	1.77	25.56
Abstractive or Mix	Pntr-Gen-Seq2Seq (See et al.)	35.86	10.22	7.60	29.69	32.06	9.04	2.15	25.16
Abstractive or Mix	Discourse summarizer (Cohan et al., 2018)	38.93	15.37	9.97	35.21	35.80	11.05	3.62	31.80
Abstractive or Mix	TLM-I+E (G,M) (Subramanian et al., 2019)	42.13	16.27	8.82	39.21	41.62	14.69	6.16	38.03
Abstractive or Mix	DANCER PEGASUS (Gidiotis and Tsoumakas, 2020)	46.34	19.97	-	42.42	45.01	17.60	-	40.56
Abstractive or Mix	PEGASUS (Zhang et al., 2020a)	45.97	20.15	-	28.25	44.21	16.95	-	25.67
Abstractive or Mix	BIGBIRD-Pegasus (Zaheer et al., 2020)	46.32	20.65	-	42.33	46.63	19.02	-	41.77
Extractive	SumBasic (Vanderwende et al., 2007)	37.15	11.36	5.42	33.43	29.47	6.95	2.36	26.30
Extractive	LexRank (Erkan and Radev, 2004)	39.19	13.89	7.27	34.59	33.85	10.73	4.54	28.99
Extractive	LSA (Steinberger and Ježek, 2004)	33.89	9.93	5.04	29.70	29.91	7.42	3.12	25.67
Extractive	Sent-CLF (Subramanian et al., 2019)	45.01	19.91	**12.13**	41.16	34.01	8.71	2.99	30.41
Extractive	Sent-PTR (Subramanian et al., 2019)	43.30	17.92	10.67	39.47	42.32	15.63	7.49	38.06
Extractive	Bert Ranker (Nogueira and Cho, 2019)	43.67	18.00	10.74	39.22	41.65	13.88	5.92	36.40
Extractive	BERTSUMEXT (Liu and Lapata, 2019b)	41.09	15.51	8.64	36.85	41.24	13.01	5.26	36.10
Extractive	BERTSUMEXT (SW) (Liu and Lapata, 2019b)	45.01	20.00	12.05	40.43	42.93	15.08	6.01	37.22
Extractive	Longformer-Ext (Beltagy et al., 2020)	43.75	17.37	10.18	39.71	45.24	16.88	8.06	40.03
Extractive	Reformer-Ext (Kitaev et al., 2020)	42.32	15.91	9.02	38.26	43.26	14.86	6.66	38.10
Extractive	G-BERT (Ours) (Grail et al., 2021)	**46.87**	**20.19**	12.11	**42.68**	**48.08**	**19.21**	**9.58**	**42.68**

Table 5.2: Summarization results on PubMed and arXiv. Except for BERT-based approaches, for Reformer-Ext and for Longformer-Ext, which we have reimplemented, the results of the baselines are taken from their associated paper as well as from Cohan et al. (2018). Bold results correspond to the best scores of extractive summarizers.

for arXiv, and $119,924/6,633/6,658$ for PubMed.

- The **Multi-News** summarization dataset has been proposed by Fabbri et al. (2019). It contains news articles associated with human-written summaries. Each summary is produced from a set of 2 to 6 source articles. They respectivelly contain $44,972$, $5,622$, $5,622$ examples in the train/validation/test sets.

- The **CNN/DailyMail** dataset contains news articles associated with short summaries. We use the splits of Hermann et al. (2015), where entities have not been anonymized. This dataset contains 287,226 training samples, 13,368 validation samples, and 11,490 test samples.

Table 5.1 presents some statistics on all these datasets. As one can note, for the scientific articles, the average number of tokens in the documents is way beyond the capabilities

of a standard transformer pre-trained with BERT.

Evaluation Metrics: We evaluate the quality of the extracted summaries using the ROUGE metric (Lin, 2004), and more particularly ROUGE-1 (overlap of unigrams), ROUGE-2 (overlap of bigrams), ROUGE-3 (overlap of trigrams) and ROUGE-L (longest common subsequence between the produced summary and the gold-standard one). On the Multi-News dataset, we repport the R-SU score which measures the overlap of skip bigrams with a max distance of four words as suggested in the original paper ((Fabbri et al., 2019)).

Label Generation: In order to train extractive summarizers, one needs annotations in the form of sentence-level binary labels. To compute such annotations, we follow the work of Kedzie et al. (2018) and label all sentences by greedily optimizing the ROUGE-1 score of the extracted summary against the gold-standard summary associated with each article. These labels are only used at training time, the evaluation of the extracted summaries being done against the gold-standard summaries provided in the datasets.

5.3.2 Baseline Models

We compare our approach to several well known published methods described below. These methods include SumBasic (Vanderwende et al., 2007), LexRank (Erkan and Radev, 2004), LSA (Steinberger and Ježek, 2004), Attn-Seq2Seq (Nallapati et al., 2016a), Pntr-Gen-Seq2Seq (See et al.) and Discourse-aware summarizer (Cohan et al., 2018). The results for these models are the ones reported in the paper (Cohan et al., 2018). We also report the results of Sent-CLF and Sent-PTR, which are hierarchical sentence pointer and classifier, TLM-I+E (G,M) a mixed extractive/generative transformer language model from Subramanian et al. (2019), BIGBIRD (Zaheer et al., 2020), PEGASUS (Zhang et al., 2020a) and DANCER (Gidiotis and Tsoumakas, 2020) which are three abstractive methods. Lastly, we developed several baseline models based on BERT, Longformer (Beltagy et al., 2020) and Reformer (Kitaev et al., 2020):

BERT Ranker: We used a BERT ranker, similar to Nogueira and Cho (2019) in which each sentence of the document is processed individually. We apply BERT on each sentence[1] and use a Sigmoid layer, the input of which consists of the [CLS] representation of the sentence, to model the probability of the sentence to be selected.

BERTSUMEXT has been introduced in Liu and Lapata (2019b). This model is an adaptation of BERT for extractive summarization. Because this model takes as input the concatenation of all the tokens of the document, it cannot scale to the arXiv and PubMed datasets. We propose two variants: the first one is to take as input only the first 800 tokens of the document, as suggested in the original paper. This solution is displayed as BERTSUMEXT in Table 5.2. The second is to apply BERTSUMEXT per sliding windows on the original document and to use, as a token representation, its representation in the window that maximizes its surrounding context. We name this sliding window implementation BERTSUMEXT (SW) in Table 5.2. For all experiments, we started with the original implementation[2] and adapted the code to build the sliding windows version. This implementation leverage *bert-base-uncased* pre-trained model and its associated hyperparameters. We use windows of width 800 with an overlap of 300 tokens between two following windows. If a sentence is in multiple windows, we select its [CLS] representation in the window that maximizes the number of surrounding tokens.

Longformer-Ext The Longformer model was introduced in Beltagy et al. (2020). It is an adaptation of the Transformer self-attention that scales to long sequences. We built the Longformer-Ext baseline from the Longformer implementation released by HuggingFace[3]. We add the same classification head as the one used in our model on top of the contextualized representation of the first token of each sentence to label them as selected or not in the summary.

We use the official *longformer-base-4096* pre-trained model trained by AllenAI[4]. This model is based on *RoBERTa-base* and its associated hyperparameters. To increase the

[1] This is possible as no sentence exceeds BERT token limitation.
[2] https://github.com/nlpyang/PreSumm
[3] https://github.com/huggingface/transformers
[4] https://github.com/allenai/longformer

maximal position embedding, we drop the pre-trained positional embedding parameters and train a novel token embedding layer to scale Longformer-Ext input up to 12294 tokens. This model computes a sliding self-attention with a window size of 512 tokens on all its 12 Transformer layers. Because of memory constraints, we use only local attentions and no global ones.

Reformer-Ext: The Reformer models was proposed Kitaev et al. (2020). It is an efficient adaptation of the original Transformer self-attention. Instead of computing the attention between each and every token of the input sequence, this attention is computed only between *similar* tokens. Tokens are considered as *similar* or not with a locality-sensitive hashing algorithm that separates them into multiple buckets. This technique is efficient in memory because the input sequence is divided into multiple buckets and in time since it allows parallel processing of the different buckets. However, there is no communication between these buckets.

We started from the HuggingFace implementation of Reformer to build Reformer-Ext baseline. We use a Reformer configuration composed of six layers of attention. We use Locality-Sensitive Hashing Attention with 128 buckets on the input sequence and Local Self-attention on chunks of 64 tokens. We use hidden states of dimension 256, a feed-forward layer of dimension 512, and 12 attention heads in Transformer encoders.

We also present the **Oracle** extractive results as an upper bound as well as the Lead baseline (which respectively select the first 3, 6, 7, 10 sentences for CNN/DailyMail, arXiv, PubMed and Multi-News datasets). Several models are reported only on CNN/DailyMail dataset and not on arXiv/Pubmed/Multi-News as they do not scale to long documents.

5.3.3 Implementation details

We run all our experiments using the Pytorch library (Paszke et al., 2019). We built our model using the "bert-base-uncase"[5] version of BERT and its implementation in the HuggingFace library (Wolf et al., 2019). Our architecture is composed of $L = 12$ propagation layers with a transformer hidden dimension of $h = 768$. The hidden dimension of the BiGRU is set to 384 and we share its parameters among all the propagation layers. The FFNN inside the propagation layers maps the output of the BiGRU of dimension 2×384 to a vector of dimension 768. The FFNN of the output layer is a binary classifier that projects the sentence representations of dimension 768 to an output of dimension 2.

We fine-tuned all models on the cross-entropy loss, for 5 epochs on 4 GPUs V100 and use Adam optimizer (Kingma and Ba, 2015) with the initial learning rate set to 3×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, no learning rate warmup and a linear decay of the learning rate. We used *Trigram Blocking* to avoid the repetition of trigrams in the extracted summaries as suggested in Paulus et al. (2018). Given the extracted summary so far, we only added candidate sentences that had no overlapping trigram with the current summary. We limit the summaries to 3 sentences for the CNN/DailyMail dataset, 6 sentences for arXiv, 7 for PubMed, and 10 for Multi-News. It corresponds to the respective average length of the gold summaries of these datasets (cf. Table 5.1).

5.3.4 Results

Our main extractive summarization results are shown in Tables 5.2, 5.4 and 5.3. On the arXiv, PubMed and Multi-News datasets, our model outperforms the baseline models on almost all of the reported metrics. Our approach manages to summarize long documents while preserving informativeness (evaluated by ROUGE-1) and fluency (evaluated by ROUGE-L) of the summaries. In addition to the previously published methods, our approach also improves over the BERT-based, Longformer-Ext and Reformer-Ext baselines we have developed. Among them, BERTSUMEXT, which focuses on a truncated version of

[5] https://github.com/google-research/bert

Model	R-1	R-2	R-L
Oracle	56.22	33.74	52.19
Lead-3	40.11	17.54	36.32
LATENT (Zhang et al., 2018)	41.05	18.77	37.54
NEUSUM (Zhou et al., 2018)	41.59	19.01	37.98
SUMO (Liu et al., 2019a)	41.00	18.40	37.20
TransformerExt (Liu and Lapata, 2019b)	40.90	18.02	37.17
MASK-LM^{global} (Chang et al., 2019)	41.2	19.1	37.6
PNBERT (Zhong et al., 2019a)	42.69	19.60	38.85
BERT-ext + RL (Bae et al., 2019)	42.76	19.87	39.11
HIBERT$_M$ (Zhang et al., 2019)	42.37	19.95	38.83
BERTSUMEXT (Liu and Lapata, 2019b)	43.25	20.24	39.63
BERTSUMEXT w/o interval embedding	43.20	20.22	39.59
BERTSUMEXT (large)	43.85	20.34	39.90
MatchSum (RoBERTa) (Zhong et al., 2020)	**44.41**	**20.86**	**40.55**
Reformer-Ext (Kitaev et al., 2020)	38.85	16.46	35.16
Longformer-Ext (Beltagy et al., 2020)	43.00	20.20	39.30
G-BERT (Ours)	42.93	19.81	39.20

Table 5.3: Comparison of ROUGE scores on CNN/DailyMail wrt extractive models. All results are taken from original papers but Reformer-Ext and Longformer-Ext which we have reimplemented.

the document, is the less effective. As documents are significantly longer than the 800 tokens limitation of this model, this result is not surprising. The sliding window adaptation of this model, that allows it to scale to long documents, is the one that achieves results that are the most comparable to ours. Our approach still outperforms this adaptation, demonstrating that summaries require to propagate information beyond a single BERT window.

On the CNN/DailyMail dataset, one can see that our model outperforms all the models that do not use pre-trained parameters. This includes several transformer-based and hierarchical models. However, while having comparable results, we do not achieve stronger performance than the current extractive state of the art from Zhong et al. (2020). This is not surprising as the majority of the CNN/DailyMail examples contains their oracle summary sentences in the first positions of the articles, as shown in the Supplementary Material, Appendix A.

Model	R-1	R-2	R-SU
Oracle	58.72	35.65	29.81
Lead-10	42.71	14.50	17.05
LexRank (Erkan and Radev, 2004)	38.44	13.10	13.50
BERTSUMEXT (SW) (Liu and Lapata, 2019b)	46.10	18.31	19.20
Reformer-Ext (Kitaev et al., 2020)	44.29	15.30	17.83
Longformer-Ext (Beltagy et al., 2020)	44.34	15.28	18.00
G-BERT (Ours) (Grail et al., 2021)	**46.90**	**18.34**	**19.92**

Table 5.4: Comparison of ROUGE scores on Multi-News dataset wrt extractive models.

5.4 Long Document Matching Experiments

5.4.1 Dataset Description

In addition, we evaluated the performance of our model on a long-to-long document matching task. It consists of citation recommendations of academic papers. Given a pair of scientific papers, the task is to predict whether one paper is a good citation for the other. This task has been proposed in Jiang et al. (2019). Because they do not release the dataset, we constructed a new similar one for these experiments. It is based on arXiv data released by Färber et al. (2018). In Appendix D, we provides more details about the contruction of this dataset. The produced dataset respectively contains $103,674$, $12,950$, 13.238 pairs of examples in the train/validation/test datasets. It is balanced between positive and negative pairs. To prevent leakage, the *"References"* section of all the papers has been removed, and all the citations are masked within the paper.

Evaluation Metrics: For long-to-long document matching experiments, we mesure the performance of the models with Accuracy, F_1-score, Precision and Recall metrics.

5.4.2 Baseline Models

The **Siamese Multi-depth Transformer-based Herarchical (SMITH)** model (Yang et al., 2020) is the current state-of-the-art approach on this task. This model extends BERT modes beyond its pre-trained maximal input length with a hierarchical structure. Similarly to our

Model	Accuracy	F_1	Precision	Recall
TF-IDF	72.59	71.09	75.25	67.37
SMITH (Yang et al., 2020)	82.14	84.20	74.55	96.71
Longformer (Beltagy et al., 2020)	**83.10**	**85.10**	**75.83**	96.9
Reformer (Kitaev et al., 2020)	79.79	81.81	72.55	93.78
G-BERT (ours) (Grail et al., 2021)	82.35	84.43	74.59	**97.24**

Table 5.5: Performance of our architecture, G-BERT, wrt other models on the long-to-long document matching dataset.

work, the input document is considered as a sequence of sentences. A first set of Transformer layers independently encodes the different sentences. From these sentence representations, a second set of Transformer layers construct the final document representation. In this work, *local* Transformers are firstly applied and *global* Transformers are stacked on the top of these. This is a major difference with our proposed architecture, where we interweave these two components allowing *local* representations to flow within the networks at every level of the architecture and not only on the latest layers. In addition, SMITH adds a specific pre-training process composed of the masked world language modeling task plus masked sentence block prediction task while our approach is only based on the masked language model. For the experiments, we used the official implementation of the SMITH model, and the pre-trained checkpoint officially released [6].

In addition, we propose a comparison with a standard **TF-IDF** baseline and also with **Reformer** and **Longformer** architectures presentded in the previous section.

5.4.3 Results

In Table 5.5, we present the performances of several models on the long-document citation recommendation dataset. In these long-to-long document matching results, we see that our model, G-BERT, outperforms TF-IDF and Reformer baselines. In terms of accuracy, it slightly underperforms the Longformer model and performs on par with SMITH

[6] https://github.com/google-research/google-research/tree/master/smith

Model	PubMed			arXiv		
	R-1	R-2	R-L	R-1	R-2	R-L
G-BERT	46.87	20.19	42.68	48.08	19.21	42.68
G-BERT-RoBERTa	46.02	19.29	41.84	47.42	18.62	42.03
G-BERT-PEGASUS	44.11	17.34	40.03	43.50	15.35	38.41
G-BERT-NoShare	46.84	20.19	42.63	48.11	19.30	42.75
G-BERT-AveragePool	45.24	18.13	40.94	45.71	17.36	40.43
G-BERT-Transformer	46.46	19.62	42.17	47.64	18.82	42.22

Table 5.6: Analysis of the influence of different key components of our proposed architecture.

architecture.

This task does not explicitly depend on sentence-level representations, contrary to extractive summarization, but on a global representation of the document. This could explain why we do not see more differences between our proposed model and other architectures. Despite this, our model stays competitive with state-of-the-art approaches on this task.

5.5 Further Analysis

We evaluate the impact of several elements of our proposed model in Table 5.6. We first study the influence of the underlying language model by considering both RoBERTa (Liu et al., 2019b) and PEGASUS (Zhang et al., 2020a) pre-trained models, respectively referred to as G-BERT-RoBERTa and G-BERT-PEGASUS. As one can see, the results show that BERT-base architecture performs best in terms of ROUGE scores on both arXiv and PubMed. One major difference between PEGASUS and BERT/RoBERTA pre-trained models is that BERT/RoBERTA are only encoders while PEGASUS is a pre-trained encoder/decoder architecture. This could explain why BERT/RoBERTA outperform PEGASUS on extractive summarization tasks. We then compare an alternative of our implementation of G-BERT in which the parameters of the BiGRU are not shared among all the propagation layers (G-BERT-NoShare) and found no clear difference with the version in

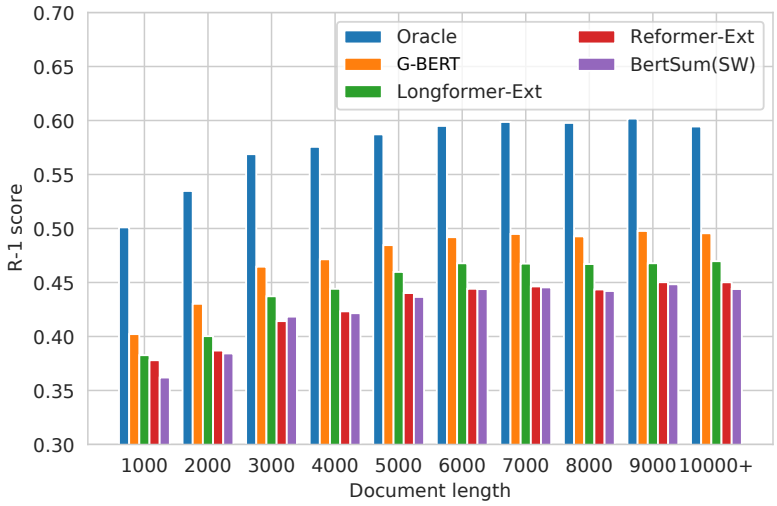

Figure 5.3: Average R-1 scores of extracted summaries according to the number of words in the input documents from arXiv test dataset.

which the parameters are shared. Lastly, we compare three architectures of propagation layers, including an average pooling of the [CLS] representations of the sentences, a Transformer layer between the [CLS] tokens (associated to a block position embedding), and a BiGRU layer. Among these three layers, the average pooling layer, which introduces no additional trainable parameters, performs the worst. Furthermore, the BiGRU layer slightly outperforms the Transformer layer in terms of ROUGE scores.

In Figure 5.3, we compare the R-1 score of several models regarding the number of words in the source documents. One can see that G-BERT consistently outperforms BERTSUMEXT (SW), Reformer-Ext and Longformer-Ext regardless of the number of words in the source documents.

We present in Table 5.7 two example summaries of a document from the PubMed test set (Kamio et al., 2009), respectively obtained by G-BERT and BERTSUMEXT (SW). The numbers in the margin indicate the position of the sentences in the original document, which is composed of a total of 78 sentences. As one can observe, G-BERT extracts sentences from various parts of the document whereas BERTSUMEXT (SW) mostly focuses on the beginning of the document. Among the sentences selected by the two models, the most

GOLD		purpose : to investigate whether the glc3a locus harboring the cyp1b1 gene is associated with normal tension glaucoma (ntg) in japanese patients.materials and methods : one hundred forty two japanese patients with ntg and 101 japanese healthy controls were recruited . patients exhibiting a comparatively early onset were selected as this suggests that genetic factors may show stronger involvement . genotyping and assessment of allelic diversity was performed on 13 highly polymorphic microsatellite markers in and around the glc3a locus.results:there were decreased frequencies of the 444 allele of d2s0416i and the 258 allele of d2s0425i in cases compared to controls (p = 0.022 and p = 0.034 , respectively) . however , this statistical significance disappeared when corrected (pc $>$ 0.05) . we did not find any significant association between the remaining 11 microsatellite markers , including d2s177 , which may be associated with cyp1b1 , and ntg (p $>$ 0.05). conclusions : our study showed no association between the glca3 locus and ntg , suggesting that the cyp1b1 gene , which is reportedly involved in a range of glaucoma phenotypes , may not be an associated factor in the pathogenesis of ntg .
GBT-EXTSUM	1-	primary open angle glaucoma (poag) is the most common type of glaucoma .
	15-	we excluded individuals who were diagnosed under 20 or over 60 years of age and who had 8.0 d or higher myopic refractive error of spherical equivalence .
	17-	the cases exhibiting a comparatively early onset were selected as they suggest that genetic factors may show stronger involvement . during diagnosis ,
	30-	the probability of association was corrected by the bonferroni inequality method , ie , by multiplying the obtained p values with the number of alleles compared .
	63-	only two adjacent markers , d2s0416i and d2s0425i , were significantly positive , as shown in table 2 , and the frequency of the 444 allele of d2s0416i and the 258 allele of d2s0425i were decreased in cases compared to controls (p = 0.022 , or = 0.59 and p = 0.034 , or = 0.42 , respectively) .
	66-	the purpose of this study was to investigate whether the glc3a locus is associated with ntg in japanese subjects , based on results from recent studies reporting that the cyp1b1 gene , located at the glc3a locus on chromosome 2p21 , could be a causative gene in poag as well as pcg . to this end , we genotyped 13 microsatellite markers in and around the glc3a locus . here
BERTSUMEXT (SW)	1-	primary open angle glaucoma (poag) is the most common type of glaucoma .
	2-	normal tension glaucoma (ntg) is an important subset of poag ; while many poag patients have high iop,1 patients with ntg have statistically normal iop.24 the prevalence of ntg is higher among the japanese population than among caucasians , and recent studies reported that 92% of poag patients in japan had ntg.58 the diagnosis of glaucoma is based on a combination of factors including optic nerve damage and specific field defects for which iop is the only treatable risk factor .
	7-	of these subjects , 142 were diagnosed with ntg , and 101 were control subjects .
	20-	genomic dna was extracted using the qiaamp dna blood mini kit (qiagen , hilden , germany) or the guanidine method . in this association study , we selected 13 highly polymorphic microsatellite markers that are located in and around the glc3a locus as shown in figure 1 .
	28-	the number of microsatellite repeats was estimated automatically using the genescan 672 software (applied biosystems) by the local southern method with a size marker of gs500 tamra (applied biosystems) .
	22-	polymerase chain reaction (pcr) was performed in a reaction mixture with a total volume of 12.5 l containing pcr buffer , genomic dna , 0.2 mm dinucleotide triphosphates (dntps) , 0.5 m primers , and 0.35 u taq polymerase .

Table 5.7: An example of summary produced by our method compared to the gold summary and one produced by BERTSUMEXT (SW). With a red scale, we highlight the sentences with the highest ROUGE score when evaluated against the abstract. We show in the margin the position of the extracted sentence in the document. This document (Kamio et al., 2009) is 78 sentences long.

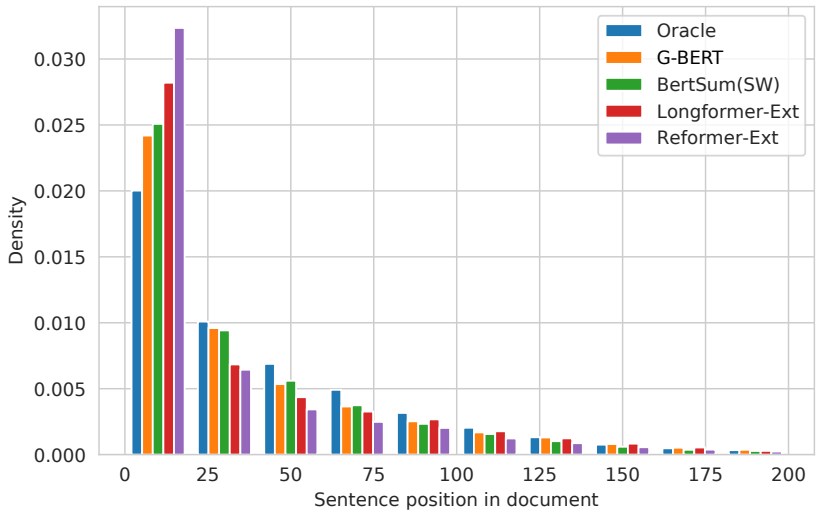

Figure 5.4: Proportion of the extracted sentences according to their position in the input document from PubMed test dataset.

meaningful one, in terms of ROUGE, is the last one selected by G-BERT. This sentence appears at position 66, in the last section (*Discussion*) of the original paper. In contrast, BERTSUMEXT (SW) proposes sentences that are less relevant for summarization purposes. Additional summaries of the PubMed and arXiv articles are provided in the Supplementary Material, Appendices B and C.

To analyse the influence of the positions of the sentences in the input document, we present in Figure 5.4 the histograms of the positions of the sentences of the Oracle summary as well as that of the predicted positions of different models, on the PubMed test set. One can see that if most relevant sentences appear at the beginning of a document, other Oracle sentences are still relevant further down the document. G-BERT is the model that behaves the most closely to the Oracle, followed by BERTSUMEXT (SW), Reformer-Ext and Longformer-Ext. These last two models tend to over-select sentences from the beginning while focusing less on the ones appearing later in the document. Our model remains influenced by the sentence position but is still able to select sentences from all over the document and is closer to the Oracle distribution.

Summary of our Contribution Related to Long Document Representations

In this part, we have introduced a novel transformer-based model for long document summarization called G-BERT. To tackle the problem of memory complexity of the original Transformer self-attention, we propose an approach based on a hierarchical architecture. The input sequence is firstly divided into multiple blocks. By doing this, it becomes possible to compute the self-attention within the blocks. In addition, we introduce propagation layers that spread information between the successive blocks. Transformer and propagation functions are interwoven in the proposed architecture, thus allowing global and local knowledge to be fused at every level of the architecture. This model preserves the architecture of commonly used pre-trained language models in order to benefit from the transfer of pre-trained parameters. An evaluation, conducted on top of the BERT model in the context of an extractive summarization task, further revealed its effectiveness in dealing with long documents compared to other adaptations of BERT and previously proposed models. Additionally, we evaluate this model on a task of long-document matching that requires predicting whether two documents are related or not by a citation link. On this former task, our model performs on par with recent scalable Transformer based architecture. Finally, an analysis of the different components of our models shows the influence of the underlying pre-trained language model and propagation layers.

Chapter 6

Conclusion and Perspectives

In this thesis, we presented our contributions related to machine reading comprehension through deep learning methods. In Chapter 2 we introduced several background requirements to guide the reader for the remaining of the thesis. We gave an overview of word embeddings models and explained how they have evolved during the last decades until the current Transformer-based models. We walked through a history of question-answering from the 1970s, presented the different formulations of the task, and illustrated them with examples from popular datasets. Eventually, we presented our proposed relational reasoning dataset ReviewQA. We concluded the background chapter with an overview of the automatic text summarization task.

Part I details our contribution related to question-answering. In Chapter 3, we introduced a reading protocol based on an adversarial couple of deep learning models. In the proposed framework, one model is in charge of answering questions while a second model is trained to introduce adversarial noise in the input document to challenge the reader. We showed that this well-chosen noise, introduced at training time, helps the reader to achieve better performance on a specific task. This protocol was proposed before the popularity of the masked language modeling task as a pre-training objective, but there is a close relationship between the two approaches that we discussed in Section 3.5. In Chapter 4, we presented our Latent Question Reformulation Network, a model based on a combination

of reading and reformulation modules. When this model was proposed, it achieved competitive results with SOTA along with interpretable attention reasoning chains as display in Figure 4.4

We see multiple possible future directions to build on these proposals. In our adversarial training protocol, we trained two models, a reader which answers questions, and an obfuscation network which corrupts the documents. In the current version of the framework, we only use the reader to produce an answer. However, we saw in several visualizations 3.4 that the obfuscation network had learned meaningful information about the task. We see two possible future directions to make use of this learned knowledge. The first one would be to integrate the word importance from the obfuscation model into the reader's input. It would help the reader to focus its attention on words that are *a-priori* labeled as important and influence it to pay less attention to the others. A drastic approach would be to directly drop words that are label as not informative from the obfuscator network point of view to simplify the input document and improve inference time and possibly reader performances. A second possible direction is to see how this obfuscation network can be transferred across tasks and datasets. At the moment, the obfuscation network is associated to a task and a dataset; however, there is no limitation to use it to obfuscate words in different contexts. It would be interesting to see if this transfer will improve the training of the reader or not affect it all.

Multi-hop was proposed as a way to tackle the limitations of mono-hop machine reading that have been highlighted in SQuAD by multiple papers. We believe that there is still a lot of progress to do in question-answering, especially with reasoning and common sense. Several datasets have been proposed recently to evaluate such competencies specifically, and we think that research in these directions will help to improve the overall understanding of machine reading models. For instance, Drop (Dua et al., 2019) proposes to evaluate discrete reasoning over paragraphs, Boratko et al. (2020) introduced ProtoQA, a new question-answering dataset for training and evaluating common sense reasoning capabilities of artificial intelligence systems. We believe that these types of benchmarks that

specifically evaluate common sense and reasoning are essential and will drive the community to focus on tasks that require a deep understanding of human language.

In the second part of this thesis, we propose G-BERT, an adaptation of the Transformer architecture that scales to long input documents. Our architecture is based on propagation layers interweaved with transformers blocks. These propagation layers allow the model to both benefit from available pre-trained parameters and scale to large context. In this work, we showed that our architecture achieves SOTA results on long document summarization.

We see multiple future research directions to build on this work. While we investigated G-BERT in the context of extractive summarization and long-documents matching, we believe that such a model could be useful for other tasks. One interesting property of this model is its ability to construct meaningful block/sentence representations at scale. These representations are contextualized in an overall document. One possible application for such representations would be in a retrieval type of approach. Indeed, having these fine-grained sentence embeddings enriched with knowledge from their source document seems promising for such tasks. The representations could also be used in the context of fact verification when formulated as a retrieval task. In a similar direction, we think another possible interest for this model is related to contextual machine translation. We know that context is essential for translation systems but computationally expensive to take into account. Such representations could be a basis for this application.

Bibliography

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sang-goo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. *CoRR*, abs/1909.08752.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Phyllis B. Baxendale. 1958. Machine-made index for technical literature - an experiment. *IBM J. Res. Dev.*, 2(4):354–361.

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

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 932–938. MIT Press.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL.

G. P. Shrivatsa Bhargav, Michael R. Glass, Dinesh Garg, Shirish K. Shevade, Saswati Dana, Dinesh Khandelwal, L. Venkata Subramaniam, and Alfio Gliozzo. 2020. Translucent answer predictions in multi-hop reading comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications*

of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 7700–7707. AAAI Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. Protoqa: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1122–1136. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2306–2317. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Claire Cardie, Vincent Ng, David R. Pierce, and Chris Buckley. 2000. Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 180–187. ACL.

Ming-Wei Chang, Kristina Toutanova, Kenton Lee, and Jacob Devlin. 2019. Language model pre-training for hierarchical document representations. *CoRR*, abs/1901.09128.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879.

Jifan Chen, Shih-Ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *CoRR*, abs/1910.02610.

Jianpeng Cheng and Mirella Lapata. 2016a. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Jianpeng Cheng and Mirella Lapata. 2016b. Neural summarization by extracting sentences

and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 93–98. The Association for Computational Linguistics.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 845–855.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 933–941, International Convention Centre, Sydney, Australia. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2694–2703, Florence, Italy. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*.

Jeffrey L. Elman. 1993. glearning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1156–1165. ACM.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of ACL 2019*.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8823–8838. Association for Computational Linguistics.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. A high-quality gold standard for citation-based tasks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2309, Florence, Italy. Association for Computational Linguistics.

D. A. Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1:1–1:15.

Kavita Ganesan. 2018. ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks. *CoRR*, abs/1803.01937.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:3029–3040.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Quentin Grail and Julien Perez. 2018. Reviewqa: a relational aspect-based opinion reading dataset. *CoRR*, abs/1810.12196.

Quentin Grail, Julien Perez, and Eric Gaussier. 2020. Latent question reformulation and information accumulation for multi-hop machine reading.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing BERT-based transformer architectures for long document summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1792–1810, Online. Association for Computational Linguistics.

Quentin Grail, Julien Perez, and Tomi SIlander. 2018. Adversarial networks for machine reading. *Traitement Automatique des Langues*, 59-02:77–100.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1311–1320, International Convention Centre, Sydney, Australia. PMLR.

Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: An automatic question-answerer. In *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*, IRE-AIEE-ACM '61 (Western), page 219–224, New York, NY, USA. Association for Computing Machinery.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Xiaoxiao Guo, Tim Klinger, Clemens Rosenbaum, Joseph P. Bigus, Murray Campbell, Ban Kawas, Kartik Talamadupula, Gerry Tesauro, and Satinder Singh. 2017. Learning to query, reason, and answer questions on ambiguous texts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *ICLR*.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999*, pages 325–332. ACL.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. Association for Computational Linguistics.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 795–806. ACM.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. 2015. Self-paced curriculum learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2694–2700. AAAI Press.

Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop reasoning. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4473–4483. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5802–5807. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918, Berlin, Germany. Association for Computational Linguistics.

M Kamio, Akira Meguro, Masao Ota, N Nomura, Kenji Kashiwagi, F Mabuchi, Hiroyuki Iijima, K Kawase, T Yamamoto, M Nakamura, Akira Negi, T Sagara, Teruo Nishida, M Inatani, Hidenobu Tanihara, M Aihara, M Araie, Takeo Fukuchi, H Abe, and Nakamura Mizuki. 2009. Investigation of the association between the glc3a locus and normal tension glaucoma in japanese patients by microsatellite analysis. *Clinical ophthalmology (Auckland, N.Z.)*, 3:183–8.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep

learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Kai A Krueger and Peter Dayan. 2009. Flexible shaping: how learning in small steps helps. *Cognition*, 110(3):380—394.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995a. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, page 68–73, New York, NY, USA. Association for Computing Machinery.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995b. A trainable document summarizer. In *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*, pages 68–73. ACM Press.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wendy Grace Lehnert. 1977. *The Process of Question Answering*. Ph.D. thesis, USA. AAI7728146.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder

for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1106–1115. The Association for Computer Linguistics.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Chengjie Sun, Zhenzhou Ji, and Bingquan Liu. 2020. Hopretriever: Retrieve hops over wikipedia to answer complex questions. *CoRR*, abs/2012.15534.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4208–4215. ijcai.org.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fei Liu and Julien Perez. 2017. Gated end-to-end memory networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5070–5081. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019b. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019a. Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

Laurens Maaten, Minmin Chen, Stephen Tyree, and Kilian Q. Weinberger. 2013. Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 410–418. JMLR Workshop and Conference Proceedings.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2020. Teacher-student curriculum learning. *IEEE Trans. Neural Networks Learn. Syst.*, 31(9):3732–3740.

Ryan T. McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, volume 4425 of *Lecture Notes in Computer Science*, pages 557–564. Springer.

Yishu Miao and Phil Blunsom. 2016. Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 319–328. The Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. A discrete hard EM approach for weakly supervised question answering. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2851–2864. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6097–6109. Association for Computational Linguistics.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2016. Virtual adversarial training for semi-supervised text classification. *CoRR*, abs/1605.07725.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016a. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016b. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *CoRR*, abs/1808.08745.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 145–152. The Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1925–1930. The Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder,

and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. 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 2553–2566, Hong Kong, China. Association for Computational Linguistics.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop QA, with evidence extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2335–2345. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David A. McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2230–2235. The Association for Computational Linguistics.

Chris D. Paice. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Inf. Process. Manag.*, 26(1):171–186.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rebecca J. Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 143–147. The Association for Computer Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8024–8035.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ethan Perez, Patrick S. H. Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8864–8880. Association for Computational Linguistics.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1756–1765. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1499–1509. Association for Computational Linguistics.

Peng Qi, Haejun Lee, Oghenetegiri ”TG” Sido, and Christopher D. Manning. 2020. Retrieve, rerank, read, then iterate: Answering open-domain questions of arbitrary complexity from text. *CoRR*, abs/2010.12527.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. 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 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. 2020. Blockwise self-attention for long document understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2555–2565. Association for Computational Linguistics.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6140–6150. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of*

the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 193–203. ACL.

Ellen Riloff and Michael Thelen. 2000. A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*.

Douglas Rohde and David C. Plaut. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72:67–109.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. 2019. Mastering atari, go, chess and shogi by planning with a learned model. *CoRR*, abs/1911.08265.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond K. Wong, and Fang Chen. 2018. A graph-theoretic summary evaluation for rouge. In *Proceedings of the 2018 Conference*

on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 762–767. Association for Computational Linguistics.

Nan Shao, Yiming Cui, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Is graph structure necessary for multi-hop question answering? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7187–7192. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 801–809, USA. Curran Associates Inc.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Josef Steinberger and Karel Ježek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *In Proc. ISIM '04*, pages 93–100.

Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher J. Pal. 2019. On extractive and abstractive neural document summarization with transformer language models. *CoRR*, abs/1909.03186.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4208–4219.

Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. 2018. Intrinsic motivation and automatic curricula via asymmetric self-play. In *International Conference on Learning Representations*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014a. Dblp:conf/emnlp/sugawaraisa18orks. *CoRR*, abs/1409.3215.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014b. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Chuanqi Tan, Furu Wei, Nan Yang, Weifeng Lv, and Ming Zhou. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. *CoRR*, abs/1706.04815.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Gerald Tesauro. 1994. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Comput.*, 6(2):215–219.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200. Association for Computational Linguistics.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9073–9080. AAAI Press.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manag.*, 43(6):1606–1618.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances*

in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 2692–2700.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. Association for Computing Machinery.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010a. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 783–792. ACM.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010b. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 783–792, New York, NY, USA. Association for Computing Machinery.

Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 618–626. ACM.

Shuohang Wang and Jing Jiang. 2017. Machine comprehension using match-lstm and answer pointer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

Tong Wang, Xingdi (Eric) Yuan, and Adam Trischler. 2017a. A joint model for question answering and question generation. In *Learning to generate natural language workshop, ICML 2017*.

W. Wang, J. Auer, R. Parasuraman, I. Zubarev, D. Brandyberry, and M. P. Harper. 2000. A question answering system developed as a project in a natural language processing course. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sytems - Volume 6*, ANLP/NAACL-ReadingComp '00, page 28–35, USA. Association for Computational Linguistics.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1705–1714.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017b. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada. Association for Computational Linguistics.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019b. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5877–5881. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics (TACL)*, 6:287–302.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

William A Woods and WOODS WA. 1977. Lunar rocks in natural english: Explorations in natural language question answering.

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2020. Answering complex open-domain questions with multi-hop dense retrieval. *CoRR*, abs/2009.12756.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference*

on Information amp; Knowledge Management, CIKM '20, page 1725–1734, New York, NY, USA. Association for Computing Machinery.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018a. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018b. Qanet: Combining local convolution with global self-attention for reading comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Xingdi Yuan, Tong Wang, Caglar Gülçehre, Alessandro Sordoni, Philip Bachman,

Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Rep4NLP@ACL*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics.

Yuyu Zhang, Ping Nie, Arun Ramamurthy, and Le Song. 2020b. DDRQA: dynamic document reranking for open-domain multi-hop question answering. *CoRR*, abs/2009.07465.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019a. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1049–1058. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, Nitish Keskar, and Richard Socher. 2019b. Coarse-grain fine-grain coattention network for multi-evidence question answering. In *International Conference on Learning Representations*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1393–1398. ACL.

Appendices

A Summarization Datasets Statistics

Figure 1 presents the distribution of the document lengths in arXiv, PubMed and CNN/DailyMail, after tokenization with pre-trained BERT-base tokenizer. It also provides the histograms of the position of the [CLS] tokens of the Oracle sentences in input documents. One can see that the three datasets contain an important number of documents longer than 512 tokens, the standard length limitation of pre-trained language models. However, one can also notice that CNN/DailyMail contains a large part of its Oracle sentences within this first window of 512 tokens. As a consequence, a model that is not able to ”read” beyond this limitation is not penalized. It is also a reason why Lead baseline is quite strong on this dataset. On the contrary, on arXiv and PubMed, one can see that a large part of Oracle sentences occurs beyond this 512 tokens window. This explains why models capable of reading long sequences are required to achieve good results on these datasets.

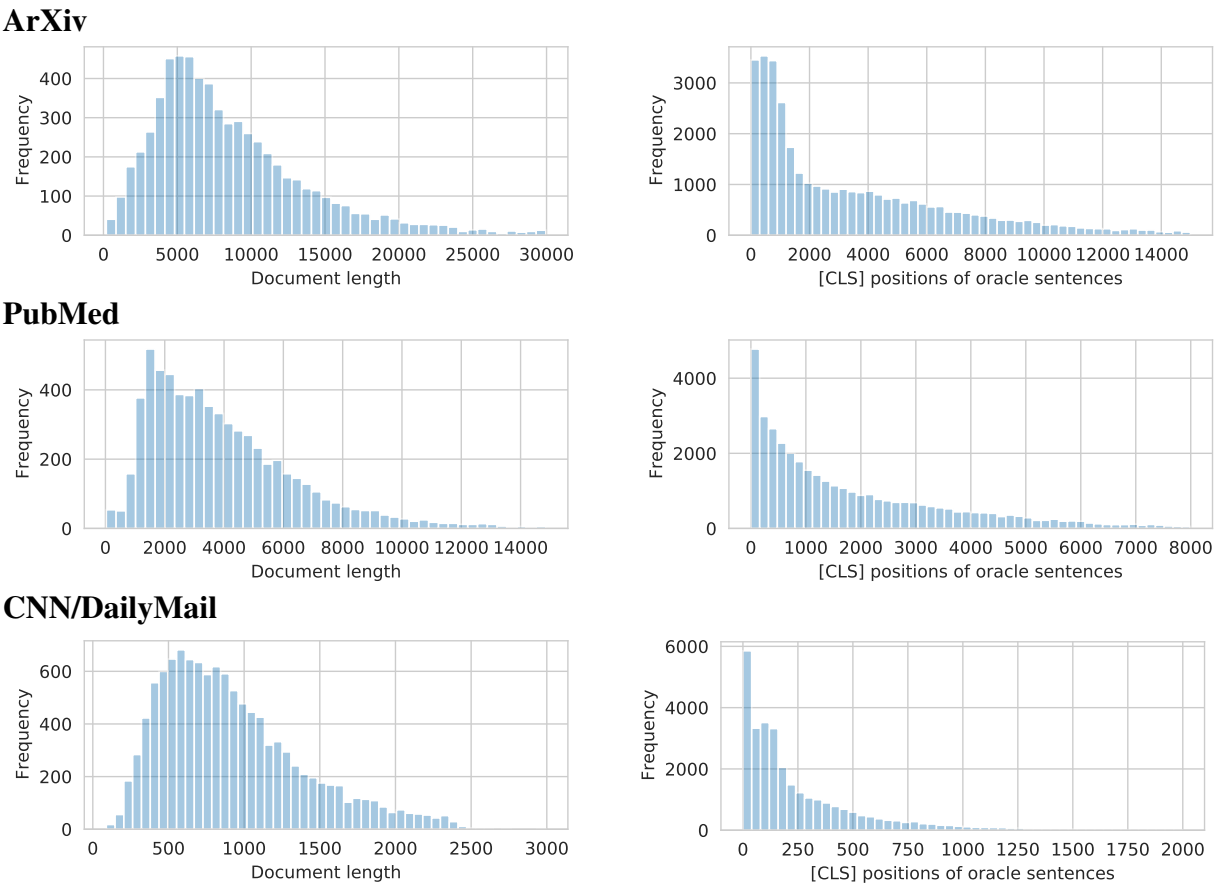

Figure 1: Document lengths after tokenization with pretrained BERT-base tokenizer and position of the [CLS] tokens of Oracle sentences in the input documents.

B PubMed Summaries

GOLD		aim . to investigate incidental adrenal enlargement clinical characteristics and functional status and analyze functional lesion risk factors . materials and methods . this retrospective study included 578 patients with adrenal imaging features showing enlargement . incidental adrenal enlargement cases (78) were considered eligible . demographics , functional diagnosis , adrenal imaging features , and concomitant diseases were analyzed . results . the number of adrenal enlargements and proportion of incidental adrenal enlargement increased each year . mean patient age was 50.32 years . thirty - nine cases had unilateral enlargement on the left side and 3 on the right side ; 36 had bilateral enlargement . routine medical checkup was found to have the greatest chance (43.59%) of revealing clinical onsets leading to discovery . biochemical and functional evaluation revealed 54 (69.23%) cases of nonfunctional lesions , 12 (15.38%) of subclinical cushing syndrome , 6 (7.69%) of primary hyperaldosteronism , 1 (1.28%) of metastasis , and 5 (6.41%) of unknown functional status . nodular adrenal enlargement (or , 7.306 ; 95% ci , 1.72728.667 ; p = 0.006) was a risk factor for functional lesions . age and lesion location were not significant factors . conclusion . incidental adrenal enlargement is a frequent radiographic finding and is accompanied by diverse clinical factors that require proper evaluation and management . nodular adrenal enlargement was a risk factor .
GBT-EXTSUM	8-	data retrieved included patient demographics , final functional diagnosis , adrenal imaging features , and concomitant diseases .
	14-	smooth enlargement was defined as enlargement of the gland with a smooth contour and no measureable or diffuse nodules . after obtaining patient history and physical examination , all patients underwent biochemical evaluation to assess their functional status .
	16-	patients with an aldosterone - rennin ratio (arr) $>$ 20 underwent any 1 of 3 confirmatory tests (saline infusion , captopril challenge , or postural stimulation) to confirm or exclude definitively primary hyperaldosteronism (pa) .
	25-	as shown in table 1 , routine medical checkup was found to have the greatest chance (43.59%) of revealing clinical onsets leading to the discovery of adrenal enlargement .
	29-	nodular adrenal enlargement (or 7.306 ; 95% ci , 1.72728.667 ; p = 0.006) was the risk factor for functional lesions .
	31-	our study shows that the proportion of incidental adrenal enlargement has gradually increased by year .
	46-	acth - independent macronodular hyperplasia (aimah) and primary pigmented nodular adrenal hyperplasia often manifest as adrenal hyperplasia . the clinical features of aimah tended to be atypical .
BERTSUMEXT (SW)	4-	it is a common term for a variety of adrenal disorders , but its cause must be properly assessed so that patients needing treatment , such as those with hormone hypersecretion or malignant disease , can receive appropriate care . however , there is a lack of literature on functional status and its follow - up to provide comprehensive insight to these findings .
	5-	patients with incidental adrenal enlargement were evaluated in a tertiary referral hospital with endocrinological departments in china .
	7-	this retrospective study included 578 patients with adrenal imaging features showing adrenal enlargement who were hospitalized at the department of endocrinology in pla general hospital (beijing , china) between january 1993 and july 2013 .
	29-	nodular adrenal enlargement (or 7.306 ; 95% ci , 1.72728.667 ; p = 0.006) was the risk factor for functional lesions .
	36-	in addition , smooth enlargement was more common , in 53 (83%) cases , and together these statistics reflect the likelihood that adrenal enlargement will be bilateral , smooth , and found in men .
	37-	however , our study did not show this tendency , likely because the research goals and thus , study populations , differed between the 2 studies .
	38-	's study aimed to explore prevalence , while the present study aimed to evaluate functional status .

GOLD		background and objective . antimicrobial resistance is now a major challenge to clinicians for treating patients . hence , this short term study was undertaken to detect the incidence of multidrug - resistant (mdr) , extensively drug - resistant (xdr) , and pandrug - resistant (pdr) bacterial isolates in a tertiary care hospital . material and methods . the clinical samples were cultured and bacterial strains were identified in the department of microbiology . the antibiotic susceptibility profile of different bacterial isolates was studied to detect mdr , xdr , and pdr bacteria . results . the antibiotic susceptibility profile of 1060 bacterial strains was studied . 393 (37.1%) bacterial strains were mdr , 146 (13.8%) strains were xdr , and no pdr was isolated . all (100%) gram negative bacterial strains were sensitive to colistin whereas all (100%) gram positive bacterial strains were sensitive to vancomycin . conclusion . close monitoring of mdr , xdr , or even pdr must be done by all clinical microbiology laboratories to implement effective measures to reduce the menace of antimicrobial resistance .
GBT-EXTSUM	5-	multidrug resistant (mdr) was defined as acquired nonsusceptibility to at least one agent in three or more antimicrobial categories . extensively drug
	36-	no mdr or xdr strain was isolated from streptococcus sp . all (100%) gram positive cocci were sensitive to vancomycin and linezolid .
	38-	e. coli was the commonest isolate 261 (35%) , followed by pseudomonas aeruginosa 212 (28.4%) .
	40-	out of 200 klebsiella pneumoniae strains isolated , 75 (37.5%) and 25 (12.5%) were detected as mdr and xdr , respectively . out of 42 acinetobacter and other nonfermenter species isolated , 19 (45.2%) and 8 (19%) were mdr and xdr strains , respectively . amongst 250 gnb - mdr strains isolated ,
	62-	, it has been reported that most frequent mdr pathogens were pseudomonas aeruginosa followed by e. coli .
	67-	unless and until multidrug resistant organisms are detected and their incidence is known , the strategies for their control can not be adopted properly in healthcare setup . hence , detection , prevention of transmission of mdros by following infection control practices , antimicrobial surveillance , and stewardship are need of the hour .
	69-	we hereby conclude that early detection and close monitoring of mdr , xdr , or even pdr bacterial strains must be started by all clinical microbiology laboratories to reduce the menace of antimicrobial resistance which is now a global problem .
BERTSUMEXT (SW)	9-	this short term cross - sectional study was conducted in the department of microbiology from 15th of april to 15th of july , 2014 .
	10-	the bacterial strains were isolated from different clinical samples and were identified by conventional methods .
	17-	methicillin resistant staphylococcus aureus (mrsa) strains were detected by meca - mediated oxacillin resistance using cefoxitin disk (30 g) on mueller hinton (mh) agar plate inoculated with test strains as per standard disk diffusion recommendations and incubated at 3335c for 1618 hours .
	20-	an increase in diameter of 5 mm with ceftazidime plus clavulanic acid as compared to ceftazidime disk alone was considered positive for esbl detection .
	36-	no mdr or xdr strain was isolated from streptococcus sp . all (100%) gram positive cocci were sensitive to vancomycin and linezolid .
	38-	e. coli was the commonest isolate 261 (35%) , followed by pseudomonas aeruginosa 212 (28.4%) .
	65-	the limitation of this study is that this is a single center study for only three - month period in a tertiary care hospital in central india . to reflect the trend of infections caused by mdr and xdr strains of bacteria in the region , a multicenter study involving all types of healthcare setups for a minimum period of one year

GOLD		background suicide is a grave public health issue that is responsible for a high mortality rate among individuals aged 1544 years . attitudes toward suicide among medical staff members have been associated with appropriate therapeutic responses to suicidal individuals . the aim of this study was to examine the effects of parental rearing on attitudes toward suicide among japanese medical college students.methodswe examined the association between parental bonding and attitudes toward suicide in 160 medical college students in japan . the parental bonding instrument was used to assess the attitudes and behaviors of parents . the attitudes toward suicide were evaluated using the japanese version of the attitudes toward suicide questionnaire.resultsthe mean age of the subjects was 25.24.0 years old . the majority of the participants in our study agreed that anyone could commit suicide (88.8%) and that suicide is preventable (86.3%) . after adjusting for age and sex , multivariate regression analysis revealed that maternal care approached a statistically significant association with the right to suicide attitude . under the same conditions , maternal care was shown to be significantly associated with the common occurrence attitude . no other significant relationships were observed between parental bonding and attitudes toward suicide.conclusionthis study suggests that a higher level of maternal care ensures that children think that suicide occurs less commonly . the promotion of best practices for suicide prevention among medical students is needed . child rearing support might be associated with suicide prevention .
GBT-EXTSUM	3-	previous studies have shown that difficulties with parental bonding during childhood could be a predisposing factor for the onset of many psychiatric conditions , such as anxiety , depressive states , and maladjusted behaviors.68 parental bonding and premorbid personality traits play an important role in shaping the developmental trajectory of an individual , including his / her ability to adjust to stressful events .
	5-	the objective of this study was to investigate whether parental bonding is associated with attitudes toward suicide among medical college students in japan .
	8-	the demographic data (age and sex) were obtained from self - questionnaires and interviews .
	14-	higher scores on the care and protection dimensions reveal that participants perceive their parents to be more caring and/or protective .
	39-	right to suicide was significantly associated with common occurrence , unjustified behavior , and preventability / readiness to help .
	43-	the majority of the participants in our study agreed that anyone could commit suicide (88.8%) and that suicide is preventable (86.3%) .
	44-	in addition , the multiple regression analysis revealed that participants who reported a higher level of maternal care thought that suicide was a common occurrence and tended to think that people do not have the right to commit suicide .
BERTSUMEXT (SW)	6-	students in their fifth year of medical school at hirosaki university , hirosaki , japan , participated in the study .
	7-	the surveys were distributed to 226 medical students . of the distributed 226 surveys , 160 questionnaires (116 males and 44 females)
	13-	the overprotection dimension of the pbi reflects parental overprotection and control in contrast to the encouragement of autonomy .
	14-	higher scores on the care and protection dimensions reveal that participants perceive their parents to be more caring and/or protective .
	15-	we employed the japanese version of the attitudes toward suicide questionnaire (atts) to assess the attitudes toward suicide held by the study participants.12 we employed a six factor model that was previously developed in studies of japanese attitudes , including
	16-	common occurrence , suicidal expression as mere threat , unjustified behavior ,
	17-	impulsiveness.12,13 each item , with the exception of items 10 and 28 , was scored on a five point scale from 1 (strongly agree) to 5 (strongly disagree) .

GOLD		introduction stasis filling , defined as delayed , weak , and persistent opacification of proximal segments of the cerebral arteries , is frequently found in brain dead patients . this phenomenon causes a major problem in the development of reliable computed tomographic angiography (cta) protocol in the diagnosis of brain death (bd) . the aim of our study was to characterize stasis filling in the diagnosis of bd . to achieve this , we performed a dynamic evaluation of contrast enhancement of the cerebral and extracranial arteries in patients with bd and controls.methodsstudy population included 30 bd patients , who showed stasis filling in computed tomographic perfusion (ctp) series . thirty patients , after clipping of an intracranial aneurysm , constituted the control group . the study protocol consisted of cta , ctp , and angiography . time density curves (tdcs) of cerebral and extracranial arteries were generated using 40-s series of ctp.resultscerebral tdcs in bd patients represented flat curves in contrast to tdcs in controls , which formed steep and narrow gaussian curves . we found longer time to peak enhancement in bd patients than in controls (32 vs. 21 s ; p ¡ 0.0001) . in bd patients , peak enhancement in the cerebral arteries occurred with a median delay of 14.5 s to peak in extracranial arteries , while no delay was noted in controls (p ¡ 0.0001) . cerebral arteries in bd patients showed lower peak enhancement than controls (34.5 vs. 81.5 hu ; p ¡ 0.0001) . in all bd patients , ctp revealed zero values of cerebral blood flow and volume . angiography showed stasis filling in 14 (46.7 %) and non - filling in 16 (53.3 %) cases.conclusiona confrontation of stasis filling with ctp results showed that stasis filling is not consistent with preserved cerebral perfusion , thus does not preclude diagnosis of bd .
GBT-EXTSUM	6-	when cta was proposed as the new imaging technique in the diagnostics of bd , a consensus on its interpretation criteria has not been reached .
	7-	an analysis of stasis filling in a dynamic series of computed tomographic perfusion (ctp) can provide valuable information on the interpretation of cta results and relation of this phenomenon to brain perfusion .
	113-	ctp detected complete absence of brain perfusion reflected by zero values of cbf and cbv in all cases .
	131-	, we found a statistically significant trend towards shorter time to peak and shorter delay of peak in the cerebral arteries of bd patients with craniectomy compared to those without it .
	140-	however , this contrast contamination should not significantly change values of delay of cerebral peak or c / e peak ratio as they were calculated on the basis of both cerebral and extracranial tcds .
	141-	in this study , we assessed the characteristic features of intracranial filling in bd patients delay and weakness of cerebrovascular opacification .
	142-	it led to the conclusion that delayed and weak opacification of cerebral arteries do not necessarily mean the presence of cerebral perfusion , thus does not preclude diagnosis of bd .
BERTSUMEXT (SW)	7-	an analysis of stasis filling in a dynamic series of computed tomographic perfusion (ctp) can provide valuable information on the interpretation of cta results and relation of this phenomenon to brain perfusion .
	8-	the aim of this prospective study was to characterize stasis filling phenomenon in the diagnosis of bd . to achieve this
	14-	the population consisted of 18 men and 12 women with a median age of 54.5 years (range , 2284 years) .
	18-	the patients were normoventilated and mean arterial blood pressure (mabp) maintained at greater than 80 mmhg .
	26-	all cta and aortocervical angiography examinations in bd patients were performed using the same methodology as described previously [4 , 5] .
	41-	this software is based on a maximum slope method , which assumes that there is no venous outflow from the tissue volume under consideration during the time of observation .
	63-	rois were automatically propagated over the entire series of scans . for each roi , time density curve (tdc) was plotted .

GOLD		excess weight has generally been associated with adverse health outcomes ; however , the link between overweight and health outcomes may vary with socioeconomic , cultural , and epidemiological conditions . we examine associations of weight with indicators of biological risk in three nationally representative populations : the us national health and nutrition examination survey , the english longitudinal study of ageing , and the social environment and biomarkers of aging study in taiwan . indicators of biological risk were compared for obese (defined using body mass index (bmi) and waist circumference) and normal weight individuals aged 54 + . generally , obesity in england was associated with elevated risk for more markers examined ; obese americans also had elevated risks except that they did not have elevated blood pressure (bp) . including waist circumference in our consideration of bmi indicated different links between obesity and waist size across countries ; we found higher physiological dysregulation among those with high waist but normal bmi compared to those with normal waist and normal bmi . americans had the highest levels of biological risk in all weight / waist groups . cross - country variation in biological risk associated with obesity may reflect differences in health behaviors , lifestyle , medication use , and culture .
GBT-EXTSUM	0-	rising levels of obesity are becoming a worldwide phenomenon and are increasingly identified as a health problem across the globe [14] .
	27-	we examine how elevated weight and obesity (using an indicator that considers both bmi and waist circumference) relate to having levels defined as clinical risk for cardiovascular , metabolic , and inflammatory markers in three aging societies that are now relatively similar in life expectancy but that differ in the timing of the epidemiological transition and obesity epidemic , history of economic development , socioeconomic levels , general lifestyle habits , health behaviors , and health care systems : the us , england , and taiwan .
	42-	we examine the following indicators of physiological dysregulation often associated with obesity and also associated with increased risk for multiple adverse health outcomes and obesity [43 , 44] : (1) cardiovascular markers : high systolic (sbp) and diastolic blood pressure (dbp) ; (2) metabolic markers : high levels of blood lipids (total and low - density lipoprotein (ldl) cholesterol , and fasting triglycerides) , low levels of high - density lipoprotein (hdl) cholesterol , and high fasting glucose and glycated hemoglobin ; (3) high levels of inflammatory markers c - reactive protein (crp ; available in nhanes and elsa) and interleukin-6 (il-6 ; available in sebas) , as crp and il-6 have been positively associated with bmi . for each indicator
	95-	england has the second highest biological risk score within each weight category , followed by taiwan . among 65-year - old women with the noted lifestyle behaviors and with a normal bmi and high waist
	98-	first , obesity is associated with physiological dysregulation in all countries with differences in the links between specific indicators of biological risk and obesity .
	134-	the increasing use of biological information to inform our understanding of health represents an innovative method in biodemography that will further contribute to the testing of current comparative theory and the potential creation of new paradigms surrounding the influence of modernization on health .
	144-	it is possible that these lifestyle behaviors are key explanatory factors to the noted cross - country differences in obesity - related biological risk .
BERTSUMEXT (SW)	2-	obesity among aging populations is relatively recent and aging among people who have been obese for much of their lives is also a new phenomenon . from 1980 to 2004 , the prevalence of obesity in the us has continued to rise from about 17% to 25% for men aged 5059 . while obesity in england has also increased during this period , from approximately 9% in 1980 to 15% in 2004 for men aged 5564 , the level of obesity remains much lower in england .
	29-	we use data from three nationally representative samples : the us national health and nutrition examination survey (nhanes , 20032006 ; n = 3855) , the english longitudinal study of ageing (elsa , 2004 - 2005 ; n = 9139) , and the social environment and biomarkers of aging study (sebas) in taiwan (2000 ; n = 1023) .
	60-	high total and ldl cholesterol is more common among the english ; lower levels of plasma glucose , crp , and glycated hemoglobin are also characteristic of the english .
	97-	this study observes three general findings about how biological risk is associated with obesity in three countries that differ in lifestyle and culture .
	102-	second , these relationships remain after controlling for demographic factors , participation in physical activity , and other behavioral factors .
	103-	third , similar to obese older adults , high waist individuals with normal bmi also exhibit greater physiological dysregulation in all countries compared to their normal bmi and normal waist counterparts .
	104-	our finding of a higher physiological dysregulation , as shown by the alternate biological risk summary score , in taiwan compared to the us and england could be due to a couple of potential explanations .

GOLD		this paper is based on linked qualitative studies of the donation of human embryos to stem cell research carried out in the united kingdom , switzerland , and china . all three studies used semi - structured interview protocols to allow an in - depth examination of donors and non - donors rationales for their donation decisions , with the aim of gaining information on contextual and other factors that play a role in donor decisions and identifying how these relate to factors that are more usually included in evaluations made by theoretical ethics . our findings have implications for one factor that has previously been suggested as being of ethical concern : the role of gratitude . our empirical work shows no evidence that interpersonal gratitude is an important factor , but it does support the existence of a solidarity - based desire to give something back to medical research . thus , we use empirical data to expand and refine the conceptual basis of bioethically theorizing the ivf stem cell interface .
GBT-EXTSUM	10-	. reproductive tissue , for example , is generally distinguished from other types of donated tissue because eggs , sperm , and embryos have the potential to give rise to new individuals , not just to prolong the lives of existing individuals , or to be used for research . because embryos are generally considered to have a different moral status from other tissues , the use of surplus embryos in research raises moral unease about the instrumentalization of human life that is not raised in quite the same way by the donation of either ova or sperm .
	17-	the authors of this paper have been involved in a series of linked qualitative studies of practices of embryo donation , first in the united kingdom (researcher haimes and colleagues) , then switzerland (scully , rehmann - sutter , and porz) , and in a smaller pilot project in china (mitzkat , rehmann - sutter , and haimes) .
	19-	design , each study was conducted independently , and the details of each project , including interview design , differed in light of the varying regulatory , clinical , and cultural contexts . however , by looking across the three data sets , we hope to gain cross - cultural insights into donation and non - donation rationales and the moral understandings on which they are based .
	30-	we then look in more detail at the implications of our findings for one area of potential ethical concern : the possible role of gratitude in making embryo disposition decisions . in this way
	31-	, we not only collect empirical data to help understand the emerging moral meaning of a new practice , but also give an example to show how empirical data can be used to question and then refine the conceptual basis of bioethical theory .
	42-	across all three studies the commonest rationale for opting to donate was a willingness to contribute to potentially curative medical research .
	156-	the reparative urge foregrounds a different set of ethical questions about the sociomoral meaning of the generation of spare embryos and the act of donation .
BERTSUMEXT (SW)	14-	understanding the social and ethical meanings that are emerging for the practices associated with embryo donation calls for a detailed empirical examination of people s reasoning behind donation decisions .
	19-	design , each study was conducted independently , and the details of each project , including interview design , differed in light of the varying regulatory , clinical , and cultural contexts . however , by looking across the three data sets , we hope to gain cross - cultural insights into donation and non - donation rationales and the moral understandings on which they are based .
	21-	interviews were designed to explore in depth not just the interviewees decision about donation but also the background to that decision , such as their ivf story , their family and other relationships , their relationship with the clinic and its staff , and so on .
	32-	participants who chose to donate their surplus embryos to research had a background premise that donation is fundamentally permissible because embryos do not have the sort of ontological or moral status that would forbid it .
	43-	such research was seen as a valuable endeavour by those like the swiss participant who said , i feel that , fundamentally , research has to go on , and i support that .
	54-	the donors were former ivf patients donating cryopreserved embryos ; in china , they were current ivf patients donating either fresh or cryopreserved embryos . in neither of these studies
	81-	note that in the chinese pilot study none of the participants spoke in precise terms about the research involved ; however , they described it broadly as scientific research for the ivf treatment .

C ArXiv Summaries

GOLD		in vivo calcium imaging through microscopes has enabled deep brain imaging of previously inaccessible neuronal populations within the brains of freely moving subjects . however , microendoscopic data suffer from high levels of background fluorescence as well as an increased potential for overlapping neuronal signals . previous methods fail in identifying neurons and demixing their temporal activity because the cellular signals are often submerged in the large fluctuating background . here we develop an efficient method to extract cellular signals with minimal influence from the background . we model the background with two realistic components : (1) one models the constant baseline and slow trends of each pixel , and (2) the other models the fast fluctuations from out - of - focus signals and is therefore constrained to have low spatial - frequency structure . this decomposition avoids cellular signals being absorbed into the background term . after subtracting the background approximated with this model , we use constrained nonnegative matrix factorization (cnmf , @xcite) to better demix neural signals and get their denoised and deconvolved temporal activity . we validate our method on simulated and experimental data , where it shows fast , reliable , and high quality signal extraction under a wide variety of imaging parameters .
GBT-EXTSUM	1-	. continued advances in optical imaging technology are greatly expanding the number and depth of neuronal populations that can be visualized .
	2-	specifically , in vivo calcium imaging through microendoscopic lenses and the development of miniaturized microscopes have enabled deep brain imaging of previously inaccessible neuronal populations of freely moving mice (@xcite) . while these techniques have been widely used by neuroscientists ,
	20-	like the proposed cnmf in @xcite , our extended cnmf for microendoscopic data (cnmf - e) also has the capability of identifying neurons with low signal - to - noise ratio (snr) and simultaneously denoising , deconvolving and demixing large - scale microendoscopic data . to accomplish this : (1) we replace the rank-1 nmf approximation of the background with a more sophisticated approximation , which can better account the complex background and avoid absorbing cellular signals , and (2) we develop an efficient initialization procedure to extract neural activities with minimal influence from the background .
	71-	@xmath56 is a template matching filter to detect spatial structures with similar shapes and sizes . for flat structures in the small regions , like background , filtering them with @xmath56
	134-	in this paper , we proposed an efficient method for extracting cellular signals from microendoscopic data ; such methods are in very high demand in the neuroscience community .
	136-	our method shows credible performances in recovering the real neuronal signals and outperforms the previous standard pca - ica method .
BERTSUMEXT (SW)	0-	monitoring the activity of large - scale neuronal ensembles during complex behavioral states is fundamental to neuroscience research
	11-	our work is based on a matrix factorization approach , which can simultaneously segment cells and estimate changes in fluorescence in the temporal domain .
	26-	the video data we have are observations from the optical field for a total number of @xmath2 frames .
	64-	we estimate the temporal component of one neuron @xmath15 from spatially filtered data and then use it to extract the corresponding spatial footprint @xmath14 from the raw data . in the step of estimating @xmath14 , we re - order all frames to make nearby frames share the similar local background levels and then take the temporal differencing to remove the background signals temporally .
	105-	we also display @xmath98 tightly clustered neurons in the simulated data (figure [fig : sim]e) to demonstrate that our cnmf - e approach can accurately detect and demix their activity (figure [fig : sim]g) .
	107-	in contrast , pca - ica based detection can only detect two neurons and the calcium traces have high level of noise .

GOLD		statistical learning theory chiefly studies restricted hypothesis classes , particularly those with finite vapnik - chervonenkis (vc) dimension . the fundamental quantity of interest is the sample complexity : the number of samples required to learn to a specified level of accuracy . here we consider learning over the set of all computable labeling functions . since the vc - dimension is infinite and a priori (uniform) bounds on the number of samples are impossible , we let the learning algorithm decide when it has seen sufficient samples to have learned . we first show that learning in this setting is indeed possible , and develop a learning algorithm . we then show , however , that bounding sample complexity independently of the distribution is impossible . notably , this impossibility is entirely due to the requirement that the learning algorithm be computable , and not due to the statistical nature of the problem .
GBT-EXTSUM	6-	an alternative approach , and one we follow in this paper , is simply to consider a single learning model that includes all possible classification methods .
	8-	since the vc - dimension is clearly infinite , there are no uniform bounds (independent of the distribution and the target concept) on the number of samples needed to learn accurately @xcite .
	10-	, it is natural to allow the learning algorithm to decide when it has seen sufficiently many labeled samples based on the training samples seen up to now and their labels . since the above learning model includes any practical classification scheme , we term it universal (pac-) learning .
	11-	we first show that there is a computable learning algorithm in our universal setting .
	19-	our results imply that computable learning algorithms in the universal setting must waste samples " in the sense of requiring more samples than is necessary for statistical reasons alone .
	81-	then we will contrast this to the case of an uncomputable learning algorithm .
BERTSUMEXT (SW)	50-	(semantic requirements) for any @xmath27 , for any concept @xmath8 , and distribution @xmath9 over @xmath2 , if the oracle returns pairs @xmath28 for @xmath29 drawn iid from @xmath9 , then @xmath0 always halts , and with probability at least @xmath12 outputs a hypothesis @xmath13 such that @xmath30 ¡ { varepsilon}$]
	64-	suppose @xmath36 is an infinite sequence of iid samples drawn from @xmath9 .
	75-	the learning algorithm queries the oracle as necessary for new learning samples and their labeling .
	78-	note that it seems necessary to expand the hypothesis space to include all partial recursive functions because the concept space of total recursive functions does not have a recursive enumeration (it is uncomputable whether a given program is total recursive or not) .
	79-	we will see in theorem [thm : nobound] that there is no bound @xmath55 on the number of samples queried by any computable learning algorithm in our setting .
	80-	let us obtain some intuition for why that is true for the above learning algorithm .

GOLD		in this paper , we propose majority voting neural networks for sparse signal recovery in binary compressed sensing . the majority voting neural network is composed of several independently trained feedforward neural networks employing the sigmoid function as an activation function . our empirical study shows that a choice of a loss function used in training processes for the network is of prime importance . we found a loss function suitable for sparse signal recovery , which includes a cross entropy - like term and an @xmath0 regularized term . from the experimental results , we observed that the majority voting neural network achieves excellent recovery performance , which is approaching the optimal performance as the number of component nets grows . the simple architecture of the majority voting neural networks would be beneficial for both software and hardware implementations .
GBT-EXTSUM	40-	requires only several matrix - vector products to obtain an output signal , which is an estimate signal of the sparse vector @xmath12 .
	48-	the signal propagates from left to right and the output signal @xmath17 eventually comes out from the output layer . the network should be trained so that the output signal @xmath17 is an accurate estimation of the original sparse signal @xmath12 .
	168-	in this paper , we proposed sparse signal recovery schemes based on neural networks for binary compressed sensing .
	169-	our empirical study shows a choice of the loss function used for training neural networks is of prime importance to achieve excellent reconstruction performance .
	170-	we found a loss function suitable for this purpose , which includes a cross entropy like term and an @xmath0 regularized term .
	173-	the simple architecture of the majority voting neural network would be beneficial for both software and hardware implementation .
BERTSUMEXT (SW)	19-	the paper @xcite presents binary iterative hard thresholding (biht) algorithm by reforming iterative hard thresholding (iht) algorithm @xcite .
	20-	although the known sparse recovery algorithms exhibit reasonable sparse recovery performance , it may not be suitable for applications in high speed wireless communications .
	48-	the signal propagates from left to right and the output signal @xmath17 eventually comes out from the output layer . the network should be trained so that the output signal @xmath17 is an accurate estimation of the original sparse signal @xmath12 .
	137-	the outputs from these neural network are combined by soft majority voting nodes and the final estimation vector is obtained by rounding the output from the soft majority voting nodes . combining a several neural networks to obtain improved performance is not a novel idea , e.g. , @xcite , but it will be shown that the idea is very effective for our purpose . from statistics of reconstruction errors occurred in our computer experiments , we observed that many reconstruction error events (i.e. , @xmath97) occur due to only one symbol mismatch .
	149-	note that implementation of neural networks with fpga is recently becoming a hot research topic @xcite .
	151-	the length of the sparse signal is set to @xmath59 and the sparseness parameter is set to @xmath110 .) , width=317] from fig.[fig : rr_and_m_k6] , we can observe significant improvement in recovery performance compared with the performance of the single neural network . a single feedforward neural network discussed in the previous section

GOLD		a path relinking algorithm is proposed for the bandwidth coloring problem and the bandwidth multicoloring problem . it combines a population based relinking method and a tabu search based local search procedure . the proposed algorithm is assessed on two sets of 66 benchmark instances commonly used in the literature . computational results demonstrate that the proposed algorithm is highly competitive in terms of both solution quality and efficiency compared to the best performing algorithms in the literature . specifically , it improves the previous best known results for 15 out of 66 instances , while matching the previous best known results for 47 cases . some key elements of the proposed algorithm are investigated . + _ keywords _ : bandwidth coloring , path relinking , tabu search , heuristics , frequency assignment .
GBT-EXTSUM	9-	for example , the fixed spectrum frequency assignment problem (fs - fap) @xcite can be viewed as a simple variant of bcp and the minimum span frequency assignment problem (ms - fap) is equivalent to bmcp considered here @xcite .
	30-	we explain in this section the ingredients of our proposed pr algorithm designed for bcp .
	134-	in this paper , we presented a pr algorithm for solving the bandwidth coloring problem and the bandwidth multicoloring problem by incorporating a tabu search algorithm with a path relinking procedure .
	136-	computational results show that our algorithm is highly competitive in comparison with the best performing algorithms in the literature . in particular , it improved best known results for 15 out of 66 instances and the improvement is very significant for several bmcp cases , yielding solutions with up to 10 fewer colors .
	137-	we studied some essential ingredients of the proposed algorithm which shed light on the following points .
	138-	first , the ts procedure is particularly appropriate as a local optimization method for our pr algorithm .
BERTSUMEXT (SW)	3-	the bandwidth multicoloring problem (bmcp) is a generalization of bcp , where each vertex @xmath5 is associated with a positive integer @xmath14 and each edge @xmath7 is associated with an edge weight @xmath9 .
	31-	bcp can be considered from the point of view of constraint satisfaction by solving a series of @xmath11-bcp problems aiming at searching for a @xmath11-coloring (@xmath11 being fixed) that satisfies all edge constraints .
	39-	specifically , let @xmath23 be a @xmath11-coloring , the objective function @xmath24 used in this study is written as : @xmath25 where @xmath9 is the edge weight for edge @xmath26 , and @xmath27 and @xmath28 respectively represent colors of vertices @xmath5 and @xmath8 .
	102-	these tests show that our pr algorithm is able to improve the best known results listed in table [results_bmcp_instances] for 14 out of 33 instances , and the improvement is impressive for some instances , such as geom120a which is solved by using 10 fewer colors than the current best solution .
	104-	, one can observe that for most instances , a smaller value of @xmath11 usually corresponds to a lower success rate and a longer average computing time for detecting a legal @xmath11-coloring .
	121-	(we used 3000 generations here .) the evolution of the best objective function value in the population and the computing time with the number of generations are separately plotted in figure [fig_alpha] for each value of @xmath112 , where the results are based on the average of 5 runs .

GOLD		we investigate two coupled nonlinear cavities that are coherently driven in a dissipative environment . we perform semiclassical , numerical and analytical quantum studies of this dimer model when both cavities are symmetrically driven . in the semiclassical analysis , we find steady - state solutions with different photon occupations in two cavities . such states can be considered analogs of the closed system double well symmetry breaking states . we analyze the occurrence and properties of these localized states in the system parameter space and examine how the symmetry breaking states , in form of a bistable pair , are associated to the single cavity bistable behavior . in a full quantum calculation of the master equation dynamics that includes quantum fluctuations , the symmetry breaking states and bistability disappear due to the quantum fluctuations . in quantum trajectory picture , we observe enhanced quantum jumps and switching which indicate the presence of the underlying semiclassical symmetry breaking states . finally , we present a set of analytical solutions for the steady state correlation functions using the complex p - representation and discuss its regime of validity .
GBT-EXTSUM	2-	thus it is an ideal system to study many outstanding questions on open systems dynamics , dissipative phase transitions @xcite and the effects of interactions in a dissipative environment .
	176-	we performed semiclassical , quantum and analytical analyses of the system . in a semiclassical treatment , we find that the nonequilibrium steady states can have asymmetric number density in the two cavities which appear in addition to the symmetry preserving states . these states are the driven - dissipative analog of the double well self - trapped or symmetry broken states .
	179-	we presented a phase diagram for these states in the tunneling - drive space .
	181-	however , in a quantum trajectory analysis of the dynamics , we found that a histogram of quantum jumps in number differences reveal the presence of semiclassical bistability with strong indication of symmetry breaking states .
	182-	finally , we presented analytical solutions for the steady state correlation functions using the complex p - representation and forming a fokker - planck equation .
	187-	the insights we gained on semiclassical and quantum nature of photons for two coupled cavities can also be useful for an array .
BERTSUMEXT (SW)	12-	we perform semiclassical and quantum analysis of the system investigating the complex interplay of many competing terms such as hopping , interaction , drive , dissipation and detuning . in a semiclassical treatment , we show that the nonequilibrium steady states have asymmetric number density in the two cavities in addition to the expected symmetry preserving states .
	29-	the cavities are coherently driven in a dissipative setting , with both drive and dissipation acting equally on both sites .
	104-	to understand more features of the driven dissipative bose - hubbard dimer , here we analyze the problem quantum mechanically taking into account the quantum fluctuations and using two methods first , we examine the dynamics by numerically solving the lindblad master equation , and second , we do a quantum trajectory or monte carlo wavefunction analysis @xcite . in fig .
	114-	we specifically would like to find out how the quantum jumps occur in the multi - stable region and whether they reveal any signature of the underlying semiclassical symmetry braking solutions .
	125-	the distribution in (f) not only shows a single peak at @xmath72 , but also broad side peaks at @xmath73 . for comparison ,
	126-	[fig : trajectory](e) and (g) show the statistics for quantum jumps outside the multistable region at @xmath74 and @xmath75 respectively , showing single lorentzian peaks at @xmath72 .

GOLD		we study the interaction between low - lying transverse collective oscillations and thermal excitations of an elongated bose - einstein condensate by means of perturbation theory . we consider a cylindrically trapped condensate and calculate the transverse elementary excitations at zero temperature by solving the linearized gross - pitaevskii equations in two dimensions . we use them to calculate the matrix elements between thermal excited states coupled with the quasi-2d collective modes . the landau damping of transverse collective modes is investigated as a function of temperature . at low temperatures , the damping rate due to the landau decay mechanism is in agreement with the experimental data for the decay of the transverse quadrupole mode , but it is too small to explain the slow experimental decay of the transverse breathing mode . the reason for this discrepancy is discussed .
GBT-EXTSUM	2-	moreover , it has been experimentally found that the transverse breathing mode of an elongated condensate exhibits unique features .
	5-	this is in contrast to the 3d case , where both monopole and quadrupole modes have similar decay rates .
	9-	we consider a cylindrical condensate and calculate at zero temperature the transverse spectrum of excitations , by solving the linearized gross - pitaevskii equation in 2d , and the modes with non - zero momentum @xmath1 along the longitudinal axis , as well .
	72-	we choose a gas of @xmath102rb atoms (scattering length @xmath103 cm) confined in an elongated cylindrical trap with frequencies @xmath104 and @xmath105 hz , that corresponds to an oscillator length @xmath106 cm .
	123-	we have investigated the decay of low - lying transverse oscillations of a large cylindrical condensate .
	126-	@xcite , we have calculated numerically the matrix elements associated with the transitions between excited states allowed by the selection rules of the transverse collective modes . within a first - order perturbation theory and assuming the thermal cloud to be in thermodynamic equilibrium , we have studied the landau damping of transverse collective modes due to the coupling with thermal excited levels as a function of temperature . for the damping rate of the transverse quadrupole mode
BERTSUMEXT (SW)	57-	the “ damping strength ” has the dimensions of a frequency and is given by @xmath78 the matrix element that couples the low - energy collective mode (@xmath79) with the higher energy single - particle excitations (for which we use the indices @xmath62) is @xcite @xmath80 . label{matrixel} end{aligned}]] in this work we calculate the quantities @xmath81 by using the numerical solutions @xmath24 and @xmath25 of eqs .
	68-	we consider the collective excitations in the collisionless regime , which is achieved at low enough temperatures . for a fixed number of trapped atoms , the number of atoms in the condensate depends on temperature .
	76-	we have solved the linearized gp equations to obtain an exact description of the ground state @xmath108 and the normal modes of the condensate within bogoliubov theory without using the thomas - fermi or hartree - fock approximations .
	78-	we have calculated the branches of the excitation spectrum of the cylindrical condensate , labeled by (@xmath49) as a function of @xmath1 . to calculate the damping rate of a collective mode @xmath112 we have to obtain , first , all pairs of levels (@xmath62) of the excitation spectrum that satisfy both the energy conservation of the transition process @xmath113 , and the selection rules given by the particular collective mode under study .
	81-	the contribution of higher excited levels can be neglected since their occupation becomes negligible in the range of temperatures we have considered .
	83-	we have checked that within our formalism the landau damping of the transverse dipole mode is zero , as expected .

GOLD		we propose a quantum feedback scheme for the preparation and protection of photon number states of light trapped in a high-@xmath0 microwave cavity . a quantum non - demolition measurement of the cavity field provides information on the photon number distribution . the feedback loop is closed by injecting into the cavity a coherent pulse adjusted to increase the probability of the target photon number . the efficiency and reliability of the closed - loop state stabilization is assessed by quantum monte - carlo simulations . we show that , in realistic experimental conditions , fock states are efficiently produced and protected against decoherence .
GBT-EXTSUM	3-	however , due to the basic quantum indetermination of the measurement outcome , measurement - induced state generation is not deterministic .
	5-	these techniques generally combine weak quantum measurements with a real - time correction of the system s state depending on the classical information extracted from the measurements . beyond preparation of specific states
	7-	, we propose a quantum feedback scheme for the on - demand preparation of fock states stored in a high - quality superconducting microwave cavity and for their protection against decoherence .
	14-	this information is , in the second step , used to estimate the new cavity field state through a quantum filtering process @xcite . in the third step
	29-	we present the quantum - mechanical operators describing the evolution of the cavity state under measurement , decoherence and pulse injection .
	218-	quantum monte - carlo simulations of the qnd measurements and the quantum feedback response demonstrate the high reliability of our closed - loop scheme even in the presence of realistic experimental imperfections .
BERTSUMEXT (SW)	48-	a second @xmath12 ramsey pulse in @xmath3 (phase @xmath20 with respect to that of the pulse in @xmath2) followed by the atomic detection (in the @xmath21 basis) by the detector @xmath4 amounts to a detection of the atomic spin along an axis at an angle @xmath20 with @xmath13 .
	51-	note that such a macroscopic time interval is well adapted to elaborate feedback strategies since we have ample time to compute the state estimator and the feedback law between two atomic detections . when no feedback action is performed , the information provided by a few tens of atoms results in a measurement of the dephasing angle @xmath15 and , hence , in a projective qnd measurement of the photon number @xmath16 @xcite .
	131-	we take into account all known imperfections of the present experimental set - up : finite cavity lifetime , poisson distribution of the atom number in atomic samples , non - ideal efficiency and state - selectivity of the detector and , finally , the finite delay between atom - cavity interaction and atomic detection .
	139-	each detection has three possible outcomes , labelled @xmath136 , @xmath137 or @xmath138 (atom detected in @xmath66 , @xmath137 or no detected atom at all) .
	143-	therefore , the state of the cavity field before injection is @xmath142 where @xmath143 includes the information gathered by the first @xmath53 sample detections and @xmath144 includes the influence of the @xmath134 in - flight samples .
	146-	this is formally taken into account by setting @xmath147 to unity for non - positive indices . if @xmath50 , i.e. for no delay in the detection process , the empty product in equals by convention the identity operator .

D Citation Recommendation Dataset Construction

In this section, we detail the workflow we used to create the dataset of long-document citation recommendations used in Section 5.4.3. We started with original data released in Färber et al. (2018). This dataset contains papers from arXiv.org released until December 31, 2017. The full dataset contains $90,278$ papers, and among them, $62,337$ are associated with a DBLP url. We only consider the papers with DBLP url as it is these urls that connect related papers together.

To construct the balanced dataset, we follow the process of Jiang et al. (2019). The *References* sections are removed to prevent leakage of ground truth. The training dataset contains pairs of related papers generated by 80% of the source papers, and validation and test sets contain related pairs of papers generated by 10% of the source papers. For each pair of related papers, we sample a negative one to create an additional pair of unrelated papers. Doing so, we obtain a balanced dataset with an equal number of related and unrelated pairs of papers. Train, validation and test datasets respectivelly contains $103,674$, $12,950$, $13,238$ pairs of papers.