Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea Magistrale in Informatica

INTEGRATING DEEP CONTEXTUALIZED WORD EMBEDDINGS INTO TEXT SUMMARIZATION SYSTEMS

Relatore:
Prof. Fabio Tamburini

Presentata da:
Claudio Mastronardo

Sessione I
Anno Accademico 2018/2019

"Considerate la vostra semenza: fatti non foste a viver come bruti ma per seguir virtute e canoscenza"

Dante Alighieri, Inferno XXVI, 116-120

Summary

In this thesis deep learning tools will be used to tackle one of the most difficult natural language processing (NLP) problems: text summarization. Given a text, the goal is to generate a summary distilling and compressing information from the whole source text. Early approaches tried to capture the meaning of text by using rules written by human. After this symbolic rule-based era, statistical approaches for NLP have taken over rule-based ones. In the last years Deep Learning (DL) has positively impacted every NLP area, including text summarization. In this work the power of pointer-generator models [See et al., 2017] is leveraged in combination with pre-trained deep contextualized word embeddings [Peters et al., 2018]. We evaluate this approach on the two largest text summarization datasets available right now: the *CNN/Daily Mail* dataset and the *Newsroom* dataset. The *CNN/Daily Mail* has been generated from the Q&A dataset published by DeepMind [Hermann et al., 2015], by concatenating sentences highlights leading to multi-sentence summaries. The Newsroom dataset is the first dataset explicitly built for text summarization [Grusky et al., 2018]. It is comprised of $\sim$1 million article-summary pairs having more or less degrees of extractiveness/abstractiveness and several compression ratios. Our approach has been evaluated on test-sets by using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score. The proposed approach leads to a great increase in performance for the Newsroom dataset achieving a state-of-the-art ROUGE-1 value and competitive values for ROUGE-2 and ROUGE-L.

Riassunto

In questa tesi saranno usate tecniche di deep learning per affrontare uno dei problemi più difficili dell'elaborazione automatica del linguaggio naturale: la generazione automatica di riassunti. Dato un corpus di testo, l'obiettivo è quello di generare un riassunto che sia in grado di distillare e comprimere l'informazione dall'intero testo di partenza. Con i primi approcci si é provato a catturare il significato del testo attraverso l'uso di regole scritte dagli umani. Dopo questa era simbolica basata su regole, gli approcchi statistici hanno preso il sopravvento. Negli ultimi anni il deep learning ha impattato positivamente ogni area dell'elaborazione automatica del linguaggio naturale, incluso la generazione automatica dei riassunti. In questo lavoro i modelli pointer-generator [See et al., 2017] sono utilizzati in combinazione a pre-trained deep contextualized word embeddings [Peters et al., 2018]. Si valuta l'approccio sui due più grossi dataset per la generazione automatica dei riassunti disponibili ora: il dataset *CNN/Daily Mail* e il dataset *Newsroom*. Il dataset *CNN/Daily Mail* è stato generato partendo dal dataset di Question Answering pubblicato da DeepMind [Hermann et al., 2015], concatenando le frasi di highlight delle news e formando così dei riassunti multi frase. Il dataset Newsroom [Grusky et al., 2018] è, invece, il primo dataset esplicitamente costruito per la generazione automatica di riassunti. Comprende un milione di coppie articolo-riassunto con diversi gradi di estrattività/astrattività a diversi ratio di compressione.

L'approccio è valutato sui test-set con l'uso della metrica Recall-Oriented Understudy for Gisting Evaluation (ROUGE). Questo approccio causa un

sostanzioso aumento nelle performance per il dataset Newsroom raggiungendo lo stato dell'arte sul valore di ROUGE-1 e valori competitivi per ROUGE-2 e ROUGE-L.

Introduction

Automatic text summarization deals with the software generation of summaries starting from a source text corpus. With the advent of internet and social media, the amount of available textual information is set to explode. In order to cope with this problem and provide fast access to information new automatic text analysis tools are needed. Automatic text summarization can constitute a powerful solution. There are mainly two types of text summarization: extractive and abstractive. Extractive text summarization techniques produce a summary by selecting words, phrases and sentences from the source article. On the other hand abstractive text summarization is capable of generating new words, new phrases and new sentences not seen in the source article. Extractive summarization while being easier and producing 100% syntactically correct summaries is considered far from a *truly* artifical intelligence based technique such as abstractive summarization, where the usage of new words and sentences demonstrates the ability of the software to understand and revise text information in order to effectively compress knowledge.

This problem has been tackled with various approaches. In this work, automatic text summarization by means of machine learning (ML) techniques is investigated. In particular, deep learning (DL) methods based on pointer-generator networks and deep contextualized word embeddings. The application of neural networks to Natural Language Processing (NLP) represents a huge step forward into the creation of more intelligent text analysis software. In the next chapters deep learning for natural language processing will be

discussed. Subsequently, the text summarization problem will be presented and formalized. Being based on machine learning, the chapter 3 will discuss the available datasets for text summarization as well as an analysis of them. The fourth chapter will present this thesis' approach as well as experimental settings and results. Considerations will be drawn at the end of the chapter. Conclusions will summarize and briefly shed light on the possible future works.

Contents

Summary i

Riassunto iv

Introduction v

1 Deep Learning for Natural Language Processing 1

1.1 Machine Learning & Deep neural networks 2

1.2 Recurrent Neural Networks 11

1.2.1 Long Short Term Memory 15

1.2.2 Gated Recurrent Unit 15

1.2.3 Bidirectional RNNs 16

1.2.4 Beam Search Decoding 18

1.3 Sequence to Sequence Models 19

1.3.1 Attention 20

1.4 Word Embeddings 24

1.4.1 Word2Vec, GloVe and fasttext 25

1.4.2 Deep Contextualized Word Embeddings 28

2 The Text Summarization Problem 33

2.0.1 Evaluation Strategies 34

2.0.2 Literature Approaches 35

3 Datasets for Text Summarization 47

3.1 Analysis of Datasets Diversity 51

4 Experiments **55**

4.1 Training results 58

Conclusions **104**

Bibliography **105**

List of Figures

1.1 Example of different representations: suppose we want to separate two categories of data by drawing a line between them in a scatterplot. In the plot on the left, we represent some data using Cartesian coordinates, and the task is impossible. In the plot on the right, we represent the data with polar coordinates and the task becomes simple to solve with a vertical line [Goodfellow et al., 2016b]. 6

1.2 A neural network and its internal data representation. It is difficult understand how the model makes its decision, but by visualizing the internal weights the network gives and each of its inputs we can see that the model utilizes raw pixels and searches for basic features such as edges, then on the knowledge extracted from basic features it builds knowledge on corners and then on object parts. Finally this rich representation of its input data is used to recognize the objects in the image. Such high dimensional and abstract way of *featurize* images would be impossible to hand craft [Goodfellow et al., 2016b]. 7

1.3 Behavior of several optimizers 10

1.4 An unfolded graph from [Goodfellow et al., 2016a] 12

1.5 An unfolded RNN[Goodfellow et al., 2016a] 13

1.6 An unfolded RNN[Goodfellow et al., 2016a] 14

1.7 A bidirectional RNN [Goodfellow et al., 2016b]. 17

1.8 Example of an encoder-decoder architecture. It is composed by the encoder that reads the input sequence and a decoder RNN that generates the output sequence. The final hidden state of the encoder is used to compute a context variable C which represents a semantic summary of the input sequence and is given to the decoder. 20

1.9 A dot-product based attention mechanism. The decoder state gets dot producted with each encoder hidden state. Each result is passed to a softmax and then a weighted sum is performed in order to give back to the decoder an *attended* context vector [Olah and Carter, 2016]. 22

1.10 Alignment matrix of a French sentence translated into English. In this matrix the white squares represent parts of the sequences where the attention weights have been high [Bahdanau et al., 2015]. 23

1.11 Starting from the string version of the word, a character embedding layer is applied to the input (for each character). The output is fed through a convolutional layer with max-pooling. Then a 2-layer highway network applies its transformations and outputs a tensor. 30

1.12 The highway network outputs the representation to a bidirectional neural network. The bidirectional neural network is executed and its hidden states are recorded. A weighted sum of the two hidden states and the input x constitute the final ELMo embedding. 31

1.13 Visualization of softmax normalized bidirectional language model network layer weights across tasks and ELMo locations. Normalized weights less then 1/3 are hatched with horizontal lines and those greater then 2/3 are speckled. Courtesy of [Peters et al., 2018] 32

2.1 Hierarchical encoder with hierarchical attention. Attention weights at the word level are re-scaled by the corresponding sentence-level attention weights. Courtesy of [Nallapati et al., 2016] 37

2.2 A visual illustration of selective encoding [Shi et al., 2018] . . 38

2.3 A visual illustration of read-again encoding [Shi et al., 2018] . 39

2.4 A visual illustration of the pointer-generator model. Courtesy of [See et al., 2017]. The encoder generates its hidden states (red). The decoder generates its hidden states (yellow). At each step the decoding process generates the attention distribution (blue), weights the hidden states and generates the context vector (red and blue). The decoder uses this information to generate the probability of generating a word from its vocabulary. Weights the generation distribution by the p_{gen} (green distribution) and sums the pointing generation after multiplying it by $1 - p_{gen}$, leading to the final distribution (on top of the image). This distribution is used by choosing the highest ranked word. 43

3.1 An overview of extractive fragment density and coverage across several only publishers used to create NEWSROOM. On the y-axis there is density and on x-axis the coverage. As shown from the plots NEWSROOM contains several types of summary generation techniques ranging from more extractive and coveraged ones such as *bostonglobe.com*, to less extractive such as *nbcnews.com*. On the top-left of each plot box n stands for the number of examples and c stands for COVERAGE. Plots are sorted by their median compression ratio. Washington post has the higher compression ratio of 27:1, while abcnews has the smallest one of 4:1. Courtesy of [Grusky et al., 2018]. . 52

3.2 Density and coverage across the top 4 datasets used in text summarization. Courtesy of [Grusky et al., 2018]. 53

4.1 Training loss curve for the newsroom dataset. Raw curve (in the background) has been smoothed with a factor of 0.99 leading to the bold line. 58
4.2 Training loss curve for the CNN/Daily Mail dataset. Raw curve (in the background) has been smoothed with a factor of 0.99 leading to the bold line. Red line shows training without coverage. Green line shows training with coverage loss enabled. Curve discontinuity (around 320k) is a normal behavior when using coverage loss. 59
4.3 Comparison of two models. The orange one represents the original pointer-generator model, the red one represents the elmo augmented pointer-generator model. 60

List of Tables

1.1 Applying ELMo embeddings to previous state-of-the-art models in six NLP tasks improves the accuracy by a margin ranging from 5.8% to 24.9%. 32

4.1 Softmax-normalized learned weights for the ELMo weighting equation. γ is the downstream task weight, while the other are per-layer weights. 60

4.2 Rouge metrics on CNN/Daily Mail dataset. This work's results are reported in bold. 61

4.3 ROUGE metric values on the Newsroom test set. This work's results are reported in bold. 61

Chapter 1

Deep Learning for Natural Language Processing

The term "language" is one of the most debated abstract concepts in terms of its definition, there exist several of such. In its most general definition it is a system consisting of the development, acquisition, maintenance and use of complex systems of communication. Focusing on human beings, we refer to as "natural language" any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation. Humans use natural languages in different forms, such as speech and signing. Languages are divided into *constructed* and *formal* ones, whereby the latter is used to program computers and study logic.
Developing systems that can understand human language is one of the main obstacles on the quest towards artificial general intelligence. This objective has driven research in artificial intelligence and particularly in natural language processing (NLP) and computational linguistics.

Early approaches towards this goal tried to capture the meaning of text using rules written by humans. Such rule-based systems, were brittle and, in general, limited to particular domains. They ultimately proved to be too restrictive to capture the intricacy of natural languages due to their lack of ability to deal with unexpected or unseen inputs; a common problem also in

other areas other than natural language processing.

After the symbolic rule-based era of such systems, a statistical approach of natural language processing has become commonplace, which uses mathematical models to automatically learn (sometimes abstract) rules from data [Manning and Schütze, 1999]. This era prompted humans into the identification of so called *features* that tell a model what connections and relationships in the data it should consider to make its prediction. The process of crafting features, however, is time-consuming as features are generally task-specific and they require domain expertise. Moreover human-crafted features represent a human-biased representation of data, which is probably not the best way to handle this type of problem.

In the past seven years, deep neural networks [Krizhevsky et al., 2017], a particular category of machine learning models, have become the model of choice when dealing with problems by using models built by learning from data. These models automatically learn a multi-layered hierarchy of features thus reducing or even removing the need for feature engineering. Human energy has transitioned to the construction of the most suitable neural network architecture and training setting for each task.

1.1 Machine Learning & Deep neural networks

Machine Learning (ML) is a sub-area of Artificial Intelligence (AI) and represents the study of designing machines (or more commonly, software for general purpose machines) that can learn from data. This is useful for solving a variety of tasks, such as computer vision and natural language processing, for which the solution is too difficult for a human software engineer to specify in terms of a fixed piece of software. A commonly-cited yet general definition states that "A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**" [Mitchell, 1997]. In this work, the experience E always includes the expe-

rience of observing a set of *examples* encoded in a specified and numerical way. For every experiment of this thesis, the experience E also includes the observation of a label for each of the examples. For classification tasks, such as emotion detection, the labels are encoded in a vector $y \in \{1, ..., k\}$, with element y_i specifying which of k object classes example i belongs to. Each numeric value in the domain of y corresponds to a real-world category such as "happy", "sad", "angry" and so on. In more complex tasks a label consists in a sequence of symbols such as words. Following this idea, the principle used to design a machine learning algorithm is called *maximum likelihood estimation* (MLE). Treating a machine learning model as a function mapping an input x to a probability using a set of *parameters* θ, as the true probability $p(x)$ of an observation is unknown, the true probability is approximated with the probability $\hat{p}_{data}(x)$ under the empirical or data generating distribution. The objective of the learning process is to bring the probability generated from the model as close as possible to the empirical probability of the input.

$$\hat{\theta}_{MLE} = argmax_{\theta}\ p_{model}(X; \theta)$$

Each task in the field of machine learning must be evaluated with a performance measure P. Depending on the nature of the task and the representation/type of labels, there exist several performance measures. In general, the evaluation phase is performed by applying the learned model to a new set of examples called *test set* and *validation set*, which represent data not seen during the learning process. Performance is obtained by taking into account the chosen measure in an objective way.

An important aspect of the performance measures described above is that they both depend on a test and validation set, this means that the learning algorithm must be able to *generalize* to new examples. Generalization is what makes machine learning different from optimization. In order to generalize one needs to assume that there exists some common structure in the data. In machine learning, there are several assumptions, one of them is the *independent and identically distributed (IID)* assumption. This assumption states that data on which we train models is comprised of mutually indepen-

dent examples and each example has the same probability distribution as the others.

Machine learning is itself divided into several classes of learning problems. Two of them are *Supervised learning* and *Unsupervised learning*.
Supervised learning is the class of learning problems where the desired output of the model on some training set is known in advance and supplied by a supervisor. One example of this is the aforementioned classification problem, where the learned model is a function $f(x)$ that maps examples x to category IDs. Another common supervised learning problem is *regression*. In the context of regression, the training set consists of a design matrix X and a vector of real-valued targets $y \in \mathbb{R}^m$.
An unsupervised learning problem is one where the learning algorithm is not provided with labels y; it is provided only with a set of examples X. The goal of an unsupervised learning algorithm is to discover something about the structure of the data.

A sub-class of machine learning representing an hybrid version between supervised and unsupervised learning is represented by *semi-supervised learning*. This class of techniques makes use of both unlabeled and labeled data for training - typically a small amount of labeled data and a large amount of unlabeled data. Since the acquisition of labeled data for a learning problem often requires a skilled human agent, the cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations semi-supervised learning can be of great practical value.

In a semi-supervised learning setting, algorithms make use of at least one of the following assumptions:

Continuity assumption: Points which are close to each other are more likely to share a label. Whereby "points" is referring to examples in the dataset and "close" is referring to a spatial proximity between them in an n-dimensional space.

Cluster assumption: The data tend to form discrete clusters, and points

in the same cluster are more likely to share a label.

Manifold assumption: The data lie approximately on a manifold of much lower dimension than the input space. In this case one can attempt to learn the manifold in order to avoid the curse of dimensionality, and possibly map high dimensional samples into a less dimensional space.

There exist several other sub-classes under the umbrella of machine learning such as dimensionality reduction, ensemble learning, meta learning, reinforcement learning so on. For the purpose of this thesis every sub-class of machine learning will not be expanded.

Feature learning (also known as *representation learning*) is an important strategy in machine learning. Many learning problems become easier if the inputs x are transformed to a new set of inputs $\phi(x)$. However, it can be difficult to explicitly design good functions ϕ. Feature learning refers to learning the feature mapping ϕ. Learned representations often result in much better performance with respect to those obtained with a hand-designed approach. They also allow AI systems to rapidly adapt to new tasks, with minimal human intervention. All of the work in this thesis employs this strategy in one way or another.

This dependence on representations is a general phenomenon that appears throughout computer science and even in daily life. In computer science, operations such as searching a collection of data can proceed exponentially faster if the collection is structured and indexed intelligently. An effective example is the one reported in [Goodfellow et al., 2016b] and in Figure 1.1.

Many feature learning algorithms are based on unsupervised learning, and can learn a reasonably useful mapping ϕ without any labeled data. This allows hybrid learning systems to improve performance on supervised learning tasks by learning features on unlabeled data. This approach constitutes the core application of this thesis. The main reason behind this approach is due to the fact that unlabeled data is usually more abundant. For example, an unsupervised learning algorithm trained on a large amount of corpus data might discover features related to the concepts of *noun* and *verb* and learn

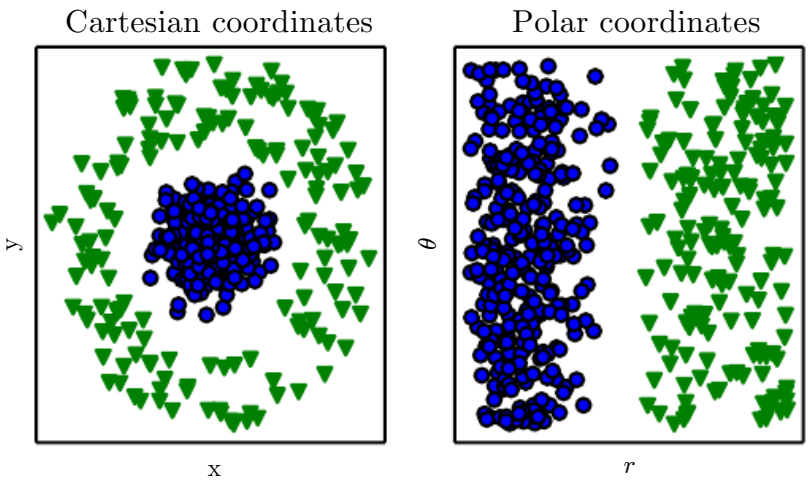

Figure 1.1: Example of different representations: suppose we want to separate two categories of data by drawing a line between them in a scatterplot. In the plot on the left, we represent some data using Cartesian coordinates, and the task is impossible. In the plot on the right, we represent the data with polar coordinates and the task becomes simple to solve with a vertical line [Goodfellow et al., 2016b].

how verbs and nouns are related in human language. A model trained on these high-level input features then needs few labeled examples in order to generalize well on a specific language-involved task.

Deep learning solves this central problem in representation learning by introducing representations that are expressed in terms of other, simpler representations. Deep learning is an approach to AI. Deep learning is a particular kind of machine learning that achieved great power and flexibility by learning to represent the world as nested hierarchy of representations, with each representation defined in relation to simpler representations, and more abstract representations computed in terms of less abstract ones. In Figure 1.2 a simple neural network and its internal input representations are reported.

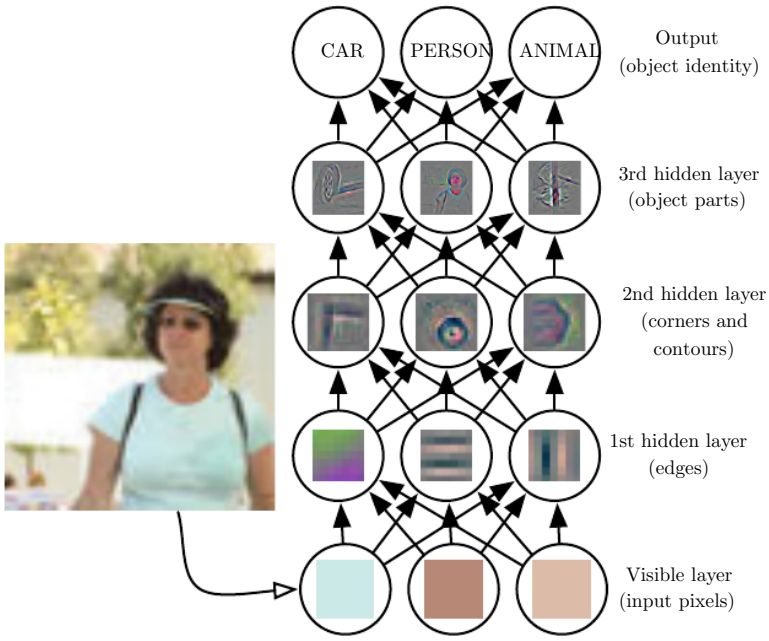

Figure 1.2: A neural network and its internal data representation. It is difficult understand how the model makes its decision, but by visualizing the internal weights the network gives and each of its inputs we can see that the model utilizes raw pixels and searches for basic features such as edges, then on the knowledge extracted from basic features it builds knowledge on corners and then on object parts. Finally this rich representation of its input data is used to recognize the objects in the image. Such high dimensional and abstract way of *featurize* images would be impossible to hand craft [Goodfellow et al., 2016b].

The whole deep learning field has been built following some rough guidelines from neuroscience. Modern deep learning draw inspiration from many fields, and the idea that many computational units become intelligent only via their interactions with each other (like in the human brain).

Neural networks can be seen as composition of functions. The standard deep learning model is the *multilayer perceptron (MLP)* also known as feed-forward neural network [Rumelhart et al., 1986]. This consists on a neural network taking some input x and using a composition of transformations

defined by several *layers*, producing an output:

$$f(x) = f_L(f_{L-1}(...f_1(x)))$$

Each layer consists of a matrix of learnable parameters $\mathbf{W}$ and a vector of learnable parameters $\mathbf{b}$ which define an affine transformation of the input. The usage of the aforementioned affine transformation as a composition of layers would constitute only of an affine transformation itself, so each layer also includes some fixed non-linear *activation function* g of the output:

$$f_i(x) = g(\mathbf{W}x + \mathbf{b})$$

where the letters in bold represent a set of learnable parameters, adjusted exploiting data in the training set. The idea of using series of layers composed of sets of units called neurons derives from the *connectionism* philosophy [Rumelhart et al., 1986] [Mehler et al., 1988]. The idea behind this philosophy is that an individual neuron in an animal or a human being is not capable of doing anything in isolation, but populations of neurons acting as a whole can achieve intelligent behavior. Similarly a single neuron of an artificial neural network is useless, but the composition of several of such, acting together, can lead to some intelligent approach of tackling a machine learning problem.[White, 1992] [Cybenko, 1989] [Hornik, 1991]
Activation functions are always chosen from the set of the non-linear functions. Example of such are the sigmoid and softmax, used for output layers, and *rectified linear unit (ReLU)* used for hidden layers.

$$ReLU(x) = max(0, x)$$

Gradient Based Learning

Parameters in a neural network are *learned* from data. The learning process is based on gradient descent, an efficient method used to minimize an objective function $J(\theta)$. It updates the model's parameters $\theta \in \mathbb{R}^d$ in the opposite direction of the *gradient* $\nabla_\theta J(\theta)$ of the function. The gradient is a vector containing all the *partial derivatives* $\frac{\partial}{\partial \theta_i} J(\theta)$. Gradient descent updates the parameters with:

$$\theta = \theta - n \times \nabla_\theta J(\theta)$$

where n is the *learning rate*, that determines the magnitude of the update of the parameters. The objective of the learning process is thus to minimize the expected value or average of an error function over the empirical distribution of the data:

$$J(\theta) = \mathbb{E}_{x,y\sim p_{data}} L(x, y, \hat{y}, \theta) = \frac{1}{n} \sum_{i=1}^{n} L(x, y, \hat{y}, \theta)$$

The gradient $\nabla_\theta J(\theta)$ is:

$$\nabla_\theta J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta L(x_i, y_i, \hat{y}_i, \theta)$$

This is known as *batch gradient descent*, which is expensive. The most used alternative is *mini-batch stochastic gradient descent* which iterates over mini-batches of m examples through the data:

$$\nabla_\theta J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta L(x_i, y_i, \hat{y}_i, \theta)$$

The mini-batch size m ranges from 2 to a few hundred depending on the training set size, typically power of 2 values.
Stochastic gradient descent (SGD) has trouble navigating ravines i.e. areas where the loss surface curves are much more steeply in one dimension than another. Momentum [Qian, 1999] is a method that helps accelerate SGD in the relevant direction by adding a fraction λ of the update vector of the past time step to the current update vector.

$$v_t = \lambda v_t + n \times \nabla_\theta J(\theta)$$
$$\theta = \theta - v_t$$

Other than the simple yet powerful gradient descent there are several other gradient based optimization algorithms [Ruder, 2016]. Some of those are more effective on some tasks more than others:

Adagrad: is an algorithm which adapts the learning rate to the parameters by performing larger updates for infrequent and smaller updates for frequent parameters.

Adadelta: is an extension of Adagrad that seeks to reduce its aggressive learning rate by restricting the window of accumulated past gradients

to some fixed size w.

RMSprop : is very similar with respect to Adadelta because it divides the learning rate by an exponentially decaying average of squared gradients.

Adam: is an extension of RMSprop and Adadelta. It keeps track of an exponentially decaying average of past gradients and uses these to update the parameters as in Adadelta.

There are also more advanced gradient based optimizers which will not be discussed in this thesis.

In Figure 1.3 an average behavior of each optimizer can be seen on the surface of the Beale function.

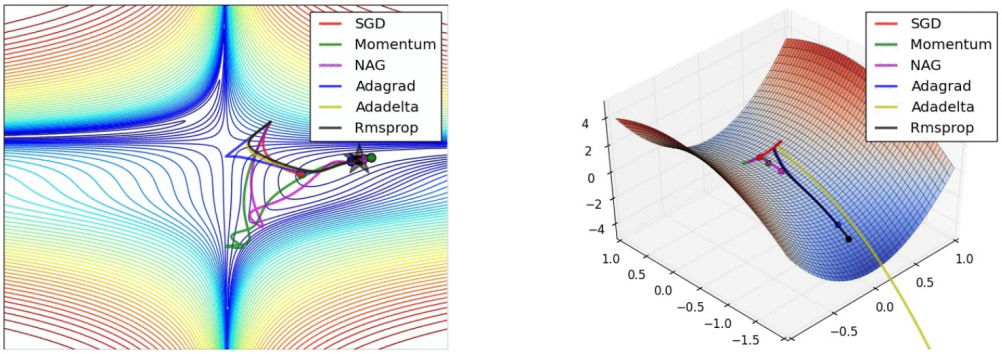

Figure 1.3: Behavior of several optimizers

Calculating the gradient of a neural network is non-trivial. In order to do so the dynamic programming algorithm known as *back-propagation* [E. Rumelhart et al., 1986] is used. It computes the gradient of a neural network by relying on the chain rule of calculus, which given functions $y = g(x)$ and $z = f(g(x)) = f(y)$ defines the derivative $\frac{dz}{dx}$ of z with respect to x as the derivative of z with respect to y times the derivative of y with respect to x. As every derivative of each layer is computed the gradient flows backward from the output layer to the input one. Each layer's derivative is a function of its successive layer's derivative. After each layer's derivative has

been computed a gradient based optimization algorithm is used to update each parameter following its layer's computed derivative.

Given a deep feed-forward neural network with L layers, weight matrices $\mathbf{W}_l$ and bias parameters $\mathbf{b}_l$ with $l \in 1, ..., L$, the model takes an input $\mathbf{x}$ and produces an output $\hat{y}$ with:

$$\mathbf{a}_l = \mathbf{b}_l + \mathbf{W}_l \mathbf{h}_{l-1}$$
$$\mathbf{h}_l = \sigma_l (\mathbf{a}_l)$$

Cost Function

There exist several cost functions which quantify the error of a machine learning for a particular inference step. For classification tasks the *cross-entropy* between the empirical conditional probability $p(y|x)$ and the probability of the model $\hat{p}(y|x;\theta)$ is used for each example x:

$$H(p, \hat{p}; x) = -\sum_{i=1}^{C} p(y_i|x) \log \hat{p}(y_i|x;\theta)$$

As a cost function $J(\theta)$, the learning process seeks to minimize the average cross-entropy over all examples in the data:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} H(p, \hat{p}; x_i)$$

1.2 Recurrent Neural Networks

The *Recurrent Neural Network* [Elman, 1990] [E. Rumelhart et al., 1986] [Goodfellow et al., 2016a] is a class of neural networks aimed at processing sequential data and they can also process sequences of variable length. The intuition behind the idea of RNNs is *parameter sharing* i.e. the idea of sharing weights across different parts of the model. This is particularly important when processing a specific piece of information that can occur at multiple positions within the sequence.

RNNs and neural networks in general, can be drawn and represented as acyclic computational directed graphs, where each node of the graph repre-

sents a transformation applied to input data. Each node outputs the results of such application and can pass this processed information to one or more nodes in the graph. To explain the repetitive structure of RNNs, the concept of *unfolding* a computation is first introduced.
Considering the classical form of a dynamical system:

$$s^{(t)} = f(s^{(t-1)}; \theta)$$

where $s^{(t)}$ is called the state of the system, one can see that this equation is recurrent because the definition of the state s at time t refers back to the same definition at time $t-1$. This is a popular way of processing time-series data and state exploration problems in artificial intelligence. For a finite number of time steps τ the graph can be unfolded by applying the definition $\tau - 1$ times. For $\tau = 3$ the definition becomes:

$$s^{(3)} = f(s^{(2)}; \theta) = f(f(s^{(2)}; \theta))$$

This new definition has been freed from the recurrent nature of the original definition and thus can be represented as a traditional directed acyclic computational graph (Figure 1.4).

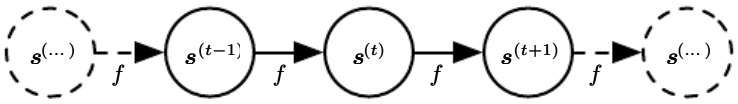

Figure 1.4: An unfolded graph from [Goodfellow et al., 2016a]

With this general way of formalizing recurrent systems the state of a RNN can be seen as its hidden units. Rewriting s with h where by h is referring to the network's hidden units:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta)$$

The sequence of inputs passed to a RNN is often called sequence of *events*. When the recurrent network is trained to perform a task that requires predicting the future from the past, it learns to use $h^{(t)}$ as a kind of lossy compressed summary of the task-relevant aspects of the past sequence of inputs up to

t. This is often regarded as the *memory* of the network. It is lossy because the hidden state of a RNN has a fixed dimension but the length of the input sequence can be variable.
An example is the statistical language modeling problem, where the task of a RNN is to predict the next word given previous words. In this task the hidden state keeps track of the essential information related to the seen words and gives the recurrent cell the ability to exploit this information in order to output a probability distribution over the set of the next possible words.

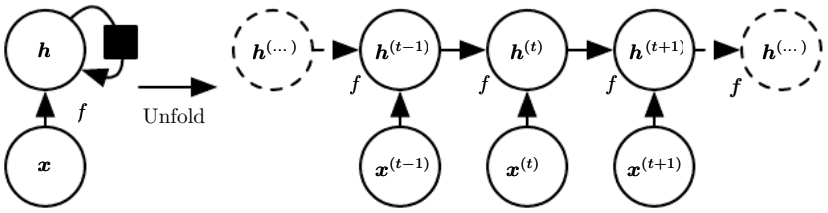

Figure 1.5: An unfolded RNN[Goodfellow et al., 2016a]

Following this setting recurrent neural networks are unfolded t times where t is the input sequence length. The unfolded recurrence after t steps can be represented with a function $g^{(t)}$:

$$h^{(t)} = g^{(t)}(x^{(t)}, x^{(t-1)}, x^{(t-2)}, ..., x^{(1)}) = f(h^{(t-1)}, x^{(t)}; \theta)$$

the function $g^{(t)}$ takes the whole past sequence $(x^{(t)}, x^{(t-1)}, x^{(t-2)}, ..., x^{(1)})$ as input and produces the current state. So regardless of the sequence length the model always has the same input size and it is possible to use the same transition function f with the same parameters. The unfolded graph provides an explicit description of which computations to perform. It also helps to illustrate the idea of information flow forward in time and backward in time (when computing the gradient) by explicitly showing the path along which this information flows. The unfolded graph is also actually used in computer hardware to perform computations.

One can use the idea of unrolling computational graphs to represent a variety of recurrent neural networks (Figure 1.6).

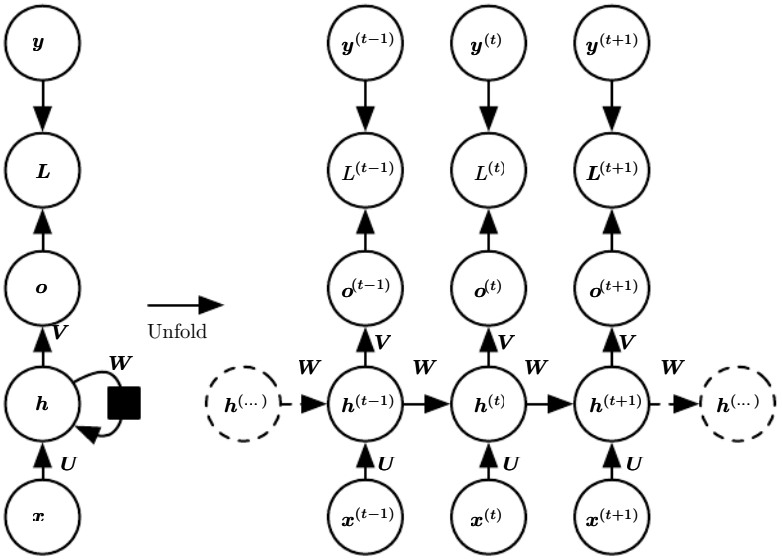

Figure 1.6: An unfolded RNN[Goodfellow et al., 2016a]

Formally speaking, a simple recurrent neural network is comprised of a recurrent cell which computes the following:

$$h_t = \sigma_h(\mathbf{W}_h x_t + \mathbf{U}_h h_{t-1} + \mathbf{b}_h)$$
$$y_t = \sigma_h(\mathbf{W}_y h_t + \mathbf{b}_y)$$

where σ_h and σ_y are activation functions. The RNNs applies a transformation U_h to modify the previous hidden state h_{t-1} and a transformation W_h to the current input x_t yielding to a new hidden state h_t. At each time step the RNN produces an output y_t.

Computing the gradient through a recurrent neural network is straightforward. One simply needs to apply the back-propagation algorithm to the unrolled computational graph of the network. The use of the back-propagation on the unrolled graph is called *back-propagation through time* (BPTT) algorithm.

RNNs struggle on handling "long-term dependencies" due to *vanishing* or *exploding* gradient problem. This phenomenon can cause the gradient to not be propagated or to exponentially increase as the RNN is run. Another type of recurrent neural network aiming to solve this problem is the *gated*

recurrent cell network.

1.2.1 Long Short Term Memory

Long Short Term Memory networks (LSTM) are one of the RNNs using gated recurrent units [Hochreiter and Schmidhuber, 1997]. This type of units allows the network to accumulate information over a long duration. Once that information has been used, it might be useful for the neural network to forget the old state. LSTM cells have an internal recurrence in addition to the outer recurrence of the RNN. The LSTM augments the RNN with a forget gate f_t, an input gate i_t and an output gate o_t, which are all function of the input x_t and the hiddent state h_t. These gates interact with the previous cell state c_{t-1}, the input and the current cell state c_t and enable the cell to selectively retain or overwrite information.

$$\begin{aligned}
\mathbf{f}_t &= \sigma_g\left(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f\right) \\
\mathbf{i}_t &= \sigma_g\left(\mathbf{W}_i x_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i\right) \\
\mathbf{o}_t &= \sigma_g\left(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o\right) \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \sigma_c\left(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c\right) \\
\mathbf{h}_t &= \mathbf{o}_t \circ \sigma_h\left(\mathbf{c}_t\right)
\end{aligned}$$

LSTMs have been shown to learn long-term dependencies more easily than the simple recurrent architectures [Sutskever et al., 2014] [Graves et al., 2013] [Graves, 2012].

1.2.2 Gated Recurrent Unit

Gated recurrent units (GRUs) aim to be a lighter version of LSTMs by having a single gating unit that simultaneously controls the forgetting factor and the decision to update the state unit. The new update equation is:

$$h_i^{(t)} = u_i^{(t-1)}h_i^{(t-1)} + (1 - u_i^{(t-1)})\sigma\left(\mathbf{b}_i + \sum_j \mathbf{U}_{i,j}x_j^{(t-1)} + \sum_j \mathbf{W}_{i,j}r_j^{(t-1)}h_j^{(t-1)}\right)$$

where u stands for the update gate and r for the reset gate. The reset

and update gates can individually ignore parts of the state vector. The update gates act like conditional leaky integrators that can linearly gate any dimension, thus choosing to copy it (at one extreme of the sigmoid) or completely ignore it (at the other extreme) by replacing it by the new "target state" value (towards which the leaky integrator wants to converge).

After several investigations over architectural variations of the LSTM and GRU, no one found better architectures on a wide range of tasks. Since the LSTM is the most used one, this thesis will use it for all of the document and experiments.

1.2.3 Bidirectional RNNs

Simple recurrent neural networks only capture information from the past i.e. $x^{(1)}, ..., x^{(t-1)}$ and the present input $x^{(t)}$. However, in many applications the output prediction $y^{(t)}$ may depend on the whole input sequence. Bidirectional recurrent neural networks [Schuster and Paliwal, 1997] have been extremely successful in applications where that need arises. Bidirectional RNNs combine an RNN that moves forward through time beginning from the start of the sequence with another RNN that moves backward through time beginning from the end of the sequence. An example is shown in Figure 1.7.

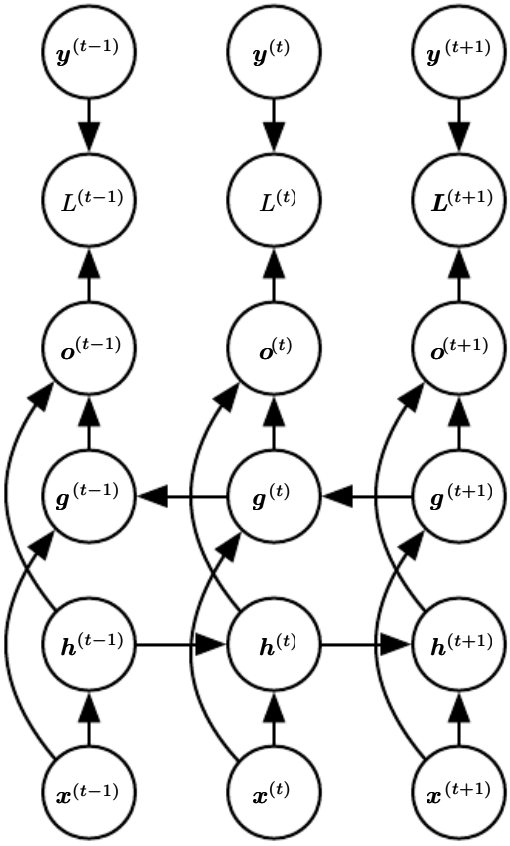

Figure 1.7: A bidirectional RNN [Goodfellow et al., 2016b].

The h recurrence propagates information forward in time while the g recurrence propagates information backward in time. Thus at each point t, the output units o can benefit from a relevant information representation that depends on both the past and the future.

The idea of bidirectional RNNs has been applied to LSTM[Graves et al., 2013] for hybrid speech recognition where the hidden state h_t is the concatenation of the hidden states from the forward and backward LSTMs at time step t:

$$h_t = [h_{fwd}; h_{bwd}]$$

1.2.4 Beam Search Decoding

A simple approach when decoding, and thus generating text with a decoder, is to select the most likely word at each step in the output sequence. This approach is undoubtedly fast, but often the quality of the final output is far from optimal. A better way to produce output is by using the popular heuristic of beam search decoding. This approach expands upon the greedy search and returns a list of most likely output sequences.

At each step instead of keeping the word with the highest probability, the beam search expands all possible next steps and keeps the k most likely, where k is a user-defined parameter and controls the size of the parallel searches. At the first step top k words are put in the beam with their corresponding output probabilities. At the next step the decoder is run on each of the k words of the beam, and produces a new distribution over the words of the vocabulary:

$$P(x_i^t|w_j^{t-1})\forall i \in 0, ..., V, \forall j \in 0, ..., k$$

where V is the size of the vocabulary and k is the size of the beam.
After each probability has been computed the top k joint probabilities are kept, and the next step is computed. At each step the beam pool contains k sequences of words which have the highest joint probabilities so far. There exist several beam search techniques for sequence to sequence models[Freitag and Al-Onaizan, 2017].

Experimentally the number k is chosen around 5 to 10. Ideally larger k lead to better performance as the multiple candidate sequences increase the likelihood of better matching a target sequence. The increased performance results in worse performance in terms of decoding speed because the model has to compute several decodings for each step of the output sequence.

1.3 Sequence to Sequence Models

Specific architectures based on recurrent neural networks can map an input sequence to an output sequence which is not necessarily of the same length. This is useful in several problems in natural language processing such as machine translation, speech recognition and text summarization, where the input and the output sequences in the training set are generally not of the same length. One of the most common approaches [Sutskever et al., 2014] [Cho et al., 2014] is comprised of an architecture that learns to encode a variable-length sequence into a fixed-length vector representation summarizing the input sequence $\mathbf{X} = (x^{(1)}, ..., x^{(n)})$. This fixed representation is decoded back into a variable-length sequence. An *encoder* RNN processes the input sequence. The encoder emits the context C, usually as a function of its final hidden state. A *decoder* RNN is conditioned on that fixed-length vector to generate the output sequence $Y = (y^{(1)}, ..., y^{n})$. The last hidden state of the encoder is the simplest way of representing the context C of the input sequence that is provided as input to the decoder RNN. The output of the decoder at each timestep t is conditioned on $\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}$ and $\mathbf{c}$; so the hidden state of the decoder at time t is:

$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}\right)$$

These two networks are jointly trained to maximize the conditional log-likelihood

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}} \left(\mathbf{y}_n | \mathbf{x}_n\right)$$

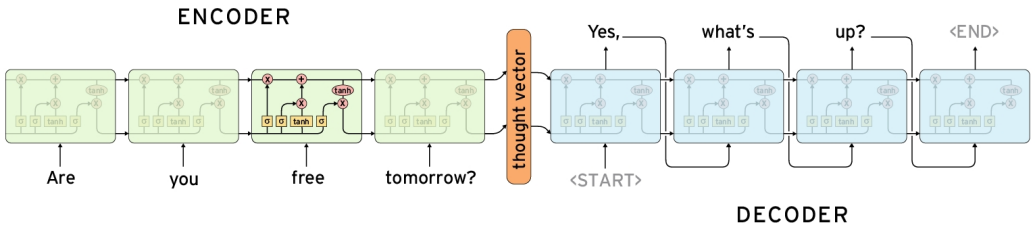

Figure 1.8: Example of an encoder-decoder architecture. It is composed by the encoder that reads the input sequence and a decoder RNN that generates the output sequence. The final hidden state of the encoder is used to compute a context variable C which represents a semantic summary of the input sequence and is given to the decoder.

A clear problem with this framework is represented by the fixed-length nature of the context C which can be too small to properly summarize a long sequence. This phenomenon has been analyzed first in the machine translation problem. [Bahdanau et al., 2015] proposed to make C a variable-length sequence rather than a fixed-size vector. Additionally, they introduced the concept of *attention mechanism* which will be now discussed.

1.3.1 Attention

[Bahdanau et al., 2015] introduced the concept of *attention* in order to overcome the problem emerged in sequence-to-sequence models where the fixed-length context vector causes the incapability of the model to remember long sentences. The solution is to create shortcuts connections where weights are customized for each output element. The alignment between source and target is learned by the context vector. Now the context vector consumes three pieces of information: encoder hidden states, decoder hidden states and some alignment between source and target sequences.

Given x as the input sequence and y as the output sequence, having n and m as length respectively:

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]$$
$$\mathbf{y} = [y_1, y_2, \ldots, y_m]$$

an encoder is a bidirectional RNN with a forward hidden state $\overrightarrow{h}_i$ and a backward hidden state $\overleftarrow{h}_i$. The encoder generates a final hidden state h_i as:

$$\boldsymbol{h}_i = \left[\overrightarrow{\boldsymbol{h}}_i^\top; \overleftarrow{\boldsymbol{h}}_i^\top\right]^\top, i = 1, \ldots, n$$

where $[a; b]$ represents the concatenation of the two vectors a and b.
The decoder network has hidden state $\boldsymbol{s}_t = f\left(\boldsymbol{s}_{t-1}, y_{t-1}, \mathbf{c}_t\right)$ for the time step t. The idea is to perform a weighted sum of the encoder hidden states:

$$\mathbf{c}_t = \sum_{i=1}^{n} \alpha_{t,t} \boldsymbol{h}_i$$

where $\alpha_{t,i}$ is:

$$\alpha_{t,i} = align(y_t, x_i) = \frac{\exp(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i))}{\sum_{i'=1}^{n} \exp(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_{i'}))}$$

which corresponds to the application of the softmax function to a vector containing *scores*. In [Bahdanau et al., 2015] the alignment score α is a feed-forward network taking the concatenation of the decoder's hidden state and each of the encoder's hidden states:

$$\text{score}\left(\boldsymbol{s}_t, \boldsymbol{h}_i\right) = \mathbf{v}_a^\top \tanh\left(\mathbf{W}_a \left[\boldsymbol{s}_t; \boldsymbol{h}_i\right]\right)$$

where both $\mathbf{v}_a$ and $\mathbf{W}_a$ are weight matrices to be learned in the alignment model. This form of attention is called *additive attention*. A more simple way of *attending* to encoder states is represented by the usage of the dot product between the decoder state and each encoder state, an example is shown in Figure 1.9.

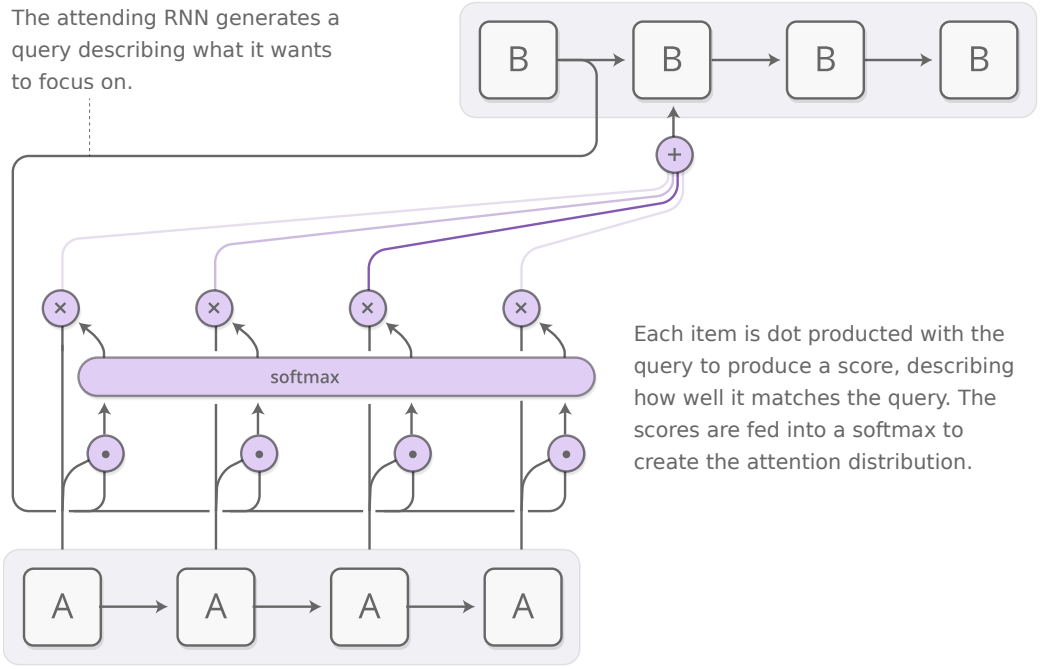

Figure 1.9: A dot-product based attention mechanism. The decoder state gets dot producted with each encoder hidden state. Each result is passed to a softmax and then a weighted sum is performed in order to give back to the decoder an *attended* context vector [Olah and Carter, 2016].

There exist several other attention mechanisms:

Content-base attention: with $\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$[Graves et al., 2014]

Location-Base: with $\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ [Luong et al., 2015]

Multiplicative: with $\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix [Luong et al., 2015]

Scaled Dot-Product: with $\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ with n as a scaling factor [Vaswani et al., 2017]

Keeping track of attention weights for each input symbol and each output symbol can lead to a better understanding of how the model is aligning source

inputs and source outputs. An example is shown in [Bahdanau et al., 2015] and reported in Figure 1.10.

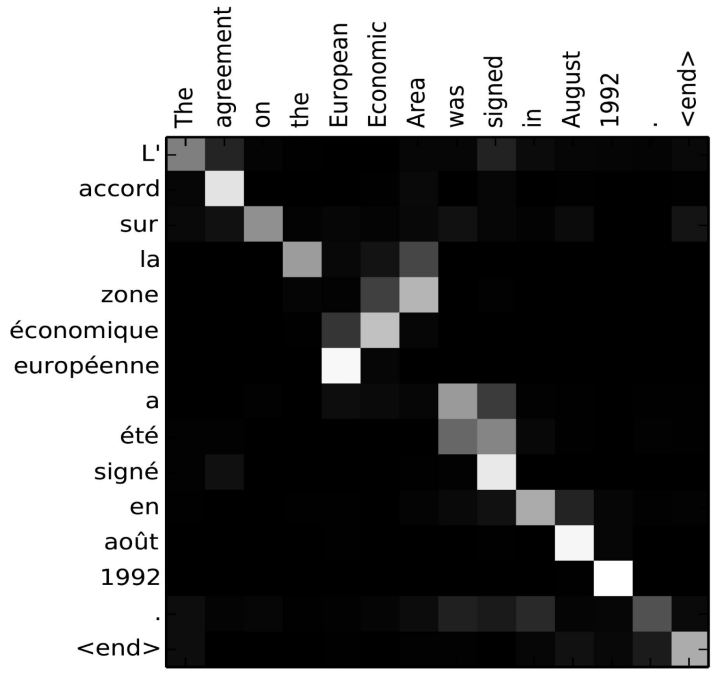

Figure 1.10: Alignment matrix of a French sentence translated into English. In this matrix the white squares represent parts of the sequences where the attention weights have been high [Bahdanau et al., 2015].

This kind of attention between RNNs has a number of other applications. It can be used in voice recognition [Chan et al., 2015] allowing one RNN to process the audio and then have another RNN skim over it, focusing on relevant parts as it generates a transcript.

The idea of attention between encoder and decoder has lead to several other attention variants such as *self-attention* and *key-value attention*. In self-attention the model is allowed to attend to itself [Lin et al., 2017]. In the recent key-value attention [Daniluk et al., 2017] each hidden state $\mathbf{h}_i$ gets split into a key $\mathbf{k}_i$ and a value $\mathbf{v}_i$: $[\mathbf{k}_i; \mathbf{v}_i] = \mathbf{h}_i$. Keys are used to calculate the attention distribution $\mathbf{a}_i$ with additive attention. The values are then used to obtain the context representation $\mathbf{c}_i$. Both $\mathbf{c}_i$ and $\mathbf{v}_i$ are used for

prediction.

1.4 Word Embeddings

Neural networks and, in general, machine learning models treat data as numbers. There have been several research proposals on how to encode actual "word strings" into numbers, in order to use statistical models. Before the deep learning revolution features were handcrafted in natural language processing too. The standard way of representing words was by encoding each word into a numerical statistics called ***tf-idf***, which stands for *term frequency - inverse document frequency*; whose aim is to reflect how important a word is to a corpus comprised of a set of *documents*. The tf-idf statistics measure multiplies the term frequency and the inverse document frequency. The term frequency simply calculates the fraction of each word's occurrences over the number of all words' occurrences. The inverse document frequency is a measure of how much information the word provides; in other terms it is encoded with the rarity of the word across all documents in a given training set. So for each word of the vocabulary a vector representing the word is created. Given a set of documents, the occurrence of each word in the vocabulary is calculated and placed in a matrix called the *term-document matrix*. This was first defined in the *vector space model idea* for information retrieval [Salton, 1971].

Following the concept of **vector semantics**, the idea shifted to representing words by vectors of real numbers. A word is represented as a point in some multidimensional semantic space. Vectors for representing words are generally called **embeddings**, because the word is embedded in a particular vector space.

Another important idea which shaped the research direction is represented by the concept that words occurring in *similar contexts* tend to have *similar meanings*, where *context* is referring to words appearing on the left and on the right side of a word in a phrase.

This link between similarity in how words are distributed and similarity in what they actually mean is captured by the **distributional hypothesis** [Joos, 1950] [Harris, 1954] [Firth, 1957].

Following this hypothesis, rather than having a term-document matrix one can create a *term-term matrix* where rows and columns are labeled by words. Each cell of the matrix now contains the number of times two words appear in the same context. Context sizes are chosen with a word window of ± 4 words around each word. The length of each word vector is now equal to the length of the vocabulary (at least 10,000). Moreover, most of these vectors are filled with zeros, making them too much **sparse**. Representing each word with a long and sparse vector is a bad idea. One can also use *hot-encoding* to represent a word. Here each word's vector is a zero-filled vector as long as the length of the vocabulary. It contains a "1" in correspondence of one dimension, and each dimension is related to one of the words of the vocabulary. But this encoding produces, again, sparse and long vectors.

In the last years alternative methods for representing words have been proposed. The common idea is to use vectors that are **short** and **dense**. They work better because a machine learning system has to learn fewer weights and thus by using fewer parameters dense vectors may generalize better and help avoid overfitting. Furthermore dense vector can do a better job in capturing synonymy because, in a typical sparse vector, words are inherently represented by distinct dimensions.

1.4.1 Word2Vec, GloVe and fasttext

In order to generate high quality, relatively short and dense word embeddings [Mikolov et al., 2013a] [Mikolov et al., 2013b] introduced the skip-gram with negative sampling algorithm, included in the software package **word2vec**.

The basic idea is to train a classifier on a binary prediction task, where the classifier is asked to predict the likelihood that a specific word w appears near a second word v. The objective is to use the learned classifier weights

as word embeddings. This approach is implicitly running a supervised task where the supervision comes to the ground truth distribution of words in every human generated corpus. This avoids the need for any sort of hand-labeled supervision signal.

A *positive example* is a couple of words where the first word is the target word and the second word is one of the words in its context. A couple with the target word and another random word is called a *negative* example. A classifier is trained to receive as input a couple and predict the probability of it being a positive or negative example. In order to produce the actual predicted value a logistic classifier is used:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Here x represents the result of the dot product of the vectors representing the words of the chosen couple, so it becomes:

$$\sigma(x) = \frac{1}{1+e^{-t \cdot c}}$$

This equation produces a probability for one word of the context. Skip-gram makes a strong assumption treating all context words as independent allowing the model to just multiply the probabilities:

$$P\left(+|t, c_{1:k}\right) = \prod_{i=1}^{k} \frac{1}{1+e^{-t \cdot c_i}}$$
$$\log P\left(+|t, c_{1:k}\right) = \sum_{i=1}^{k} \log \frac{1}{1+e^{-t \cdot c_i}}$$

In summary, skip-gram trains a probabilistic classifier that, given a test target word t and its context window of k words $c_{1:k}$, assigns a probability based on how similar the context window is to the target word.
Word2vec learns embeddings by starting with an initial set of embedding vectors and then iteratively tuning the embedding of each word to be more like the embeddings of words that occur nearby in texts.

Skip-gram uses *negative sampling*, i.e. the use of more negative samples than positive samples, where each negative sample consists of the target word t and a noise word. Noise words are drawn from a distribution where each word's probability is proportional to the number of occurrences in the dataset of each word, multiplied by a weight $\alpha = 0.75$, giving rare words a slightly

higher probability.
So given a set of positive and negative instances the goal is to adjust random initialized embeddings in order to:

- Maximize the similarity of the target word and its context words
- Minimize the similarity between the target word and other noise words drawn from the negative samples

leading the loss function to:

$$L(\theta) = \sum_{(t,c)\in+} \log P(+|t,c) + \sum_{(t,c)\in-} \log P(-|t,c)$$

maximizing the dot product of the positive words and minimizing the dot product of the negative words. The classifier is trained with stochastic gradient descent to maximize the objective.

Skip-gram learns two separate embeddings for each word: the **target embeddings** and the **context embeddings**. These embeddings are stored in two matrices, the **target matrix T** and the **context embedding C**. Each row corresponds to a word in the vocabulary. So the learning algorithm starts with randomly initialized matrices **T** and **C** and uses gradient descent to change these values.

Once embeddings are learned, each word has two embedding. One can choose to use only one embedding, some of them or even concatenate them. A powerful characteristic of these approach is that one can choose the length of the embeddings, leading the algorithm to learn more or less accurate word representations.

A very curious semantic property of these embeddings is their ability to capture relational meanings. For example [Mikolov et al., 2013a] and [Levy and Goldberg, 2014] showed that offsets between vector embeddings can capture some analogical relations between words. For example, the result of expression vector('king') - vector('man') + vector('woman') generates a vector very close to vector('queen').

There are many other embedding algorithms such as GloVe and fasttext. GloVe [Pennington et al., 2014] is based on capturing global corpus statistics

based on ratios of probabilities from word-word co-occurence matrix, combining the intuitions of count-based models while also capturing the linear structures used by word2vec. Fasttext [Bojanowski et al., 2016] deals with unknown words and sparsity by using subword models. Here each word is represented as a bag of constituent n-grams and surrounding each word with the symbols '<' and '>'. Then a skip-gram embedding is learned for each constituent n-gram.

1.4.2 Deep Contextualized Word Embeddings

Word embeddings models such as word2vec and GloVe gave a powerful boost to the natural language ecosystem in terms of quality of produced machine learning models. However, the previous techniques learned a *fixed* embedding for each word in the vocabulary. So this type of embedding is not modeling complex characteristics of the word nor how these vary across linguistic contexts.

In their word "Deep contextualized word representations" [Peters et al., 2018] propose a new way of embedding a word where the representation is a function of the entire input sentence. These vectors are derived from a bidirectional LSTM trained with a language model objective on a large text corpus. These embeddings are called ELMo (Embeddings from Language Models). ELMo embeddings are function of all of the internal layers of the bidirectional language model network. ELMo representations capture both context-dependent aspects of word meanings and syntax related aspects.

Given a sequence of N tokens $(t_1, t_2, ..., t_N)$ a forward language model computes the probability of the sequence by modeling the probability of a word conditioned on the previous words. Elmo uses a bidirectional LSTM preceded by a character level convolutional neural network. The input to the convolutional neural network is a sequence of characters which make up the word. There is no more need to list all vocabulary words and discard the unknown words, the convolutional network can apply its transformations to every conceivable sequence of characters and output a new representation.

This representation is fed into a bidirectional LSTM computing the next word's probability.
Elmo is a task specific combination of the intermediate layer representation of this LSTM. For each sequence token t_k, a L-layer bidirectional language modeling network computes a set of $2L + 1$ representations:

$$\begin{aligned} R_k &= \left\{ \mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} | j = 1, \ldots, L \right\} \\ &= \left\{ \mathbf{h}_{k,j}^{LM} | j = 0, \ldots, L \right\} \end{aligned}$$

where $\mathbf{h}_{k,0}^{LM}$ is the token layer and $\mathbf{h}_{k,j}^{LM} = \left[\overrightarrow{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM} \right]$ for each layer.
In order to include ELMo in a downstream model, one needs to collapse all these representations into a single vector:

$$\mathbf{ELMo}_k = E\left(R_k; \mathbf{\Theta}_e\right)$$

In the simplest case Elmo selects just the top layer. More generally one can compute a task specific weighting of all layers:

$$\mathbf{ELMo}_k^{\text{task}} = E\left(R_k; \Theta^{\text{task}}\right) = \gamma^{\text{task}} \sum_{j=0}^{L} s_j^{\text{task}} \mathbf{h}_{k,j}^{LM}$$

where s^{task} are softmax-normalized weights and γ^{task} allows it to scale the entire produced vector with respect to the downstream task. Both s^{task} and γ^{task} are parameters learned during training.
So, in order to utilize ELMo for supervised natural language processing tasks, the bidirectional language model network is executed and all of the layer representations for each word are recorded. Then the end task model learns a linear combination of these representations.

The pre-trained bidirectional language model used in ELMo draws its inspiration from the work of [Józefowicz et al., 2016] by using the *CNN-BIG-LSTM* with 2 layers composed of 4096 units, 512 dimension projections and a residual connection from the first to second layer. The context insensitive type representation uses 2048 character n-gram convolutional filters followed by two highway layers [Srivastava et al., 2015]. Running the convolutional layers on character embeddings allows it to pick up morphological features that simple word-level embeddings could miss. In addition, one can form

valid embeddings even for out-of-vocabulary words.
So, the bidirectional language model network provides three layers of representations for each token.
The first stage of the process is reported in Figure 1.11 [Józefowicz et al., 2016].
After this process the representation goes to a bidirectional RNN (Figure

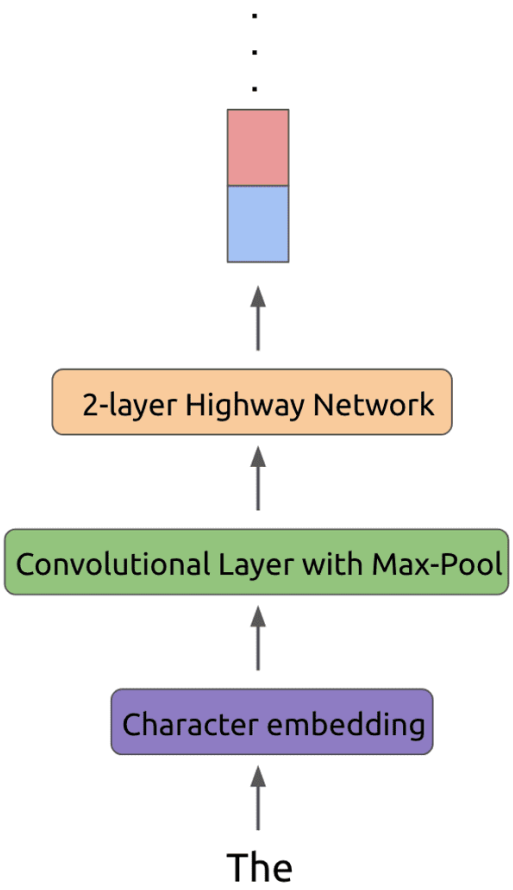

Figure 1.11: Starting from the string version of the word, a character embedding layer is applied to the input (for each character). The output is fed through a convolutional layer with max-pooling. Then a 2-layer highway network applies its transformations and outputs a tensor.

1.12).

ELMo has been applied in a variety of benchmark NLP tasks yielding to state-of-the-art results with error reductions ranging from 6% to 20% (Table

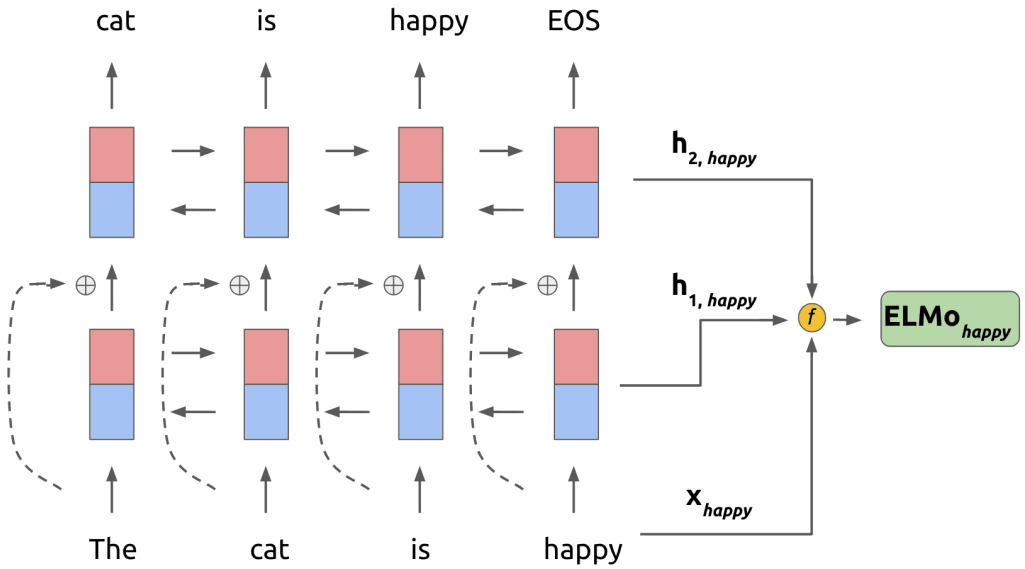

Figure 1.12: The highway network outputs the representation to a bidirectional neural network. The bidirectional neural network is executed and its hidden states are recorded. A weighted sum of the two hidden states and the input x constitute the final ELMo embedding.

1.1). The analysis of ELMo embeddings shows that syntactic information is captured at lower layers while semantic information is captured at higher layers, like machine translation encoders. Depending on the downstream task the model chooses how much weight give to lower and higher layer representations (Figure 1.13).

Several other works in this direction produced very remarkable results leading to several other state-of-the-art approaches on a variety of natural language processing problems [Peters et al., 2017] [Howard and Ruder, 2018] [Radford, 2018] [Vaswani et al., 2017].

Task	Previous SOTA		Author baseline	elmo+ baseline	Increase (absolute/ relative)
SQuAD	[Liu et al., 2017]	84.4	81.1	85.8	4.7 / 24.9%
SNLI	[Chen et al., 2016]	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	[He et al., 2017]	81.7	81.4	84.6	3.2 / 17.2%
Coref	[Lee et al., 2017]	67.2	67.2	70.4	3.2 / 9.8%
NER	[Peters et al., 2017]	91.93	90.15	92.22± 0.10	2.06 / 21%
SST-5	[McCann et al., 2017]	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Table 1.1: Applying ELMo embeddings to previous state-of-the-art models in six NLP tasks improves the accuracy by a margin ranging from 5.8% to 24.9%.

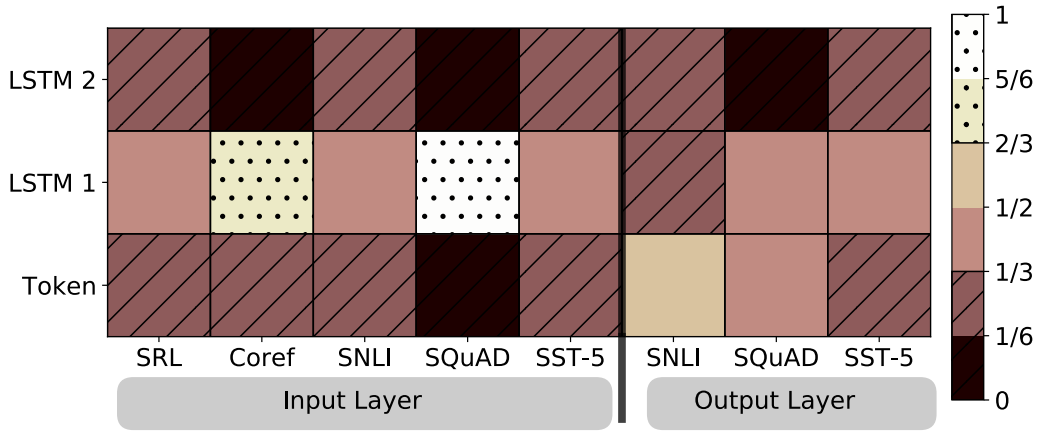

Figure 1.13: Visualization of softmax normalized bidirectional language model network layer weights across tasks and ELMo locations. Normalized weights less then 1/3 are hatched with horizontal lines and those greater then 2/3 are speckled. Courtesy of [Peters et al., 2018]

Chapter 2

The Text Summarization Problem

Among the sundry natural language processing problems, this thesis tackles one of the most difficult one: text summarization. With the advent of the big data era retrieving useful information from large sources of textual documents is a challenging task. Automatic text summarization provides an effective solution for summarizing these textual documents. The task of the text summarization is to condense long documents into short summaries while preserving the important information and meaning of the documents [Radev et al., 2002] [Allahyari et al., 2017]. According to [Radev et al., 2002] a *summary* is defined as a text that is produced from one or more texts, that conveys important information in the original text, and that is no longer than half of the original text and usually, significantly less than half". Text summarization is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and the write a summary highlighting its main points, a tough task for a computer lacking human knowledge and language capability.

Generally speaking, there are two ways to do text summarization: Extractive and Abstractive [Mani, 1999].
A method is considered *extractive* if words, phrases and sentences in the sum-

mary are completely selected from the source articles. They are relatively simple and can produce grammatically correct phrases. On the other hand, abstractive text summarization is capable of generating novel sentences using language generation models grounded on representation of source documents. They have a strong potential of producing high-quality summaries by also incorporating external knowledge.

This problem has been tackled with a variety of approaches, but in the last 5 years many deep neural network based models have achieved remarkable results in terms of the evaluation measures such us BLEU[Papineni et al., 2002] and ROUGE[Lin, 2004], explained in the next section.

2.0.1 Evaluation Strategies

There have been several evaluation approaches proposed in the literature. In order to be able to do automatic summary evaluation there are quite a few difficulties. It is fundamental to decide and specify the most important parts of the original text; these important parts need to be identified in the summary and, in general, the readability of the summary in terms of grammaticality and coherence has to be evaluated. One way of tackling the problem of evaluation is represented by human evaluation. Here, human judges read each summary and evaluate how much of the candidate summary coverages the original source document. Judges must give scores based on summaries' grammaticality, non redundancy, integration of most important pieces of information, structure and coherence. This way of evaluating summaries is obviously biased by the judge's experience and its way of reading and comprehending text.

Since early 2000s there has been a set of metric to automatically evaluate summaries; ROUGE is the most widely used one. ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation [Lin, 2004], is a metric to automatically determine the quality of a summary by comparing it to human reference summaries. ROUGE has several variations, but the most used ones are:

- **ROUGE-n**: compares n-grams. A series of 2/3/4-grams are extracted from the reference summary and the candidate summary.
  The ROUGE-n is the number of common n-grams between candidate and reference summary divided by the number of n-grams extracted from the reference summary only.

- **ROUGE-L**: this measure calculates the *longest common subsequence (LCS)* between summary sentences. So all n-grams must be consecutive.

- **ROUGE-SU**: considers bi-grams and uni-grams. It allows the insertion of words between the first and the last words of the bi-grams.

BLEU (Bilingual Evaluation Understudy) [Papineni et al., 2002] is an evaluation algorithm predominantly used in machine translation. BLEU is based on the degree of n-gram overlapping between the strings of words produced by the systems and humans. It computes the precision for n-gram of size 1-to-4 with the coefficient of brevity penalty. It is sometimes used to evaluate summarization models, but not too often.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Banerjee and Lavie, 2005] is based on the concept of flexible unigram matching, including match of words that are simple morphological variants of each other by matching their stemmed versions and words which are synonyms of each other. Other than a modified f-measure, METEOR introduces a penalty coefficient by using the number of matched chunks.

$$\text{Penalty} = 0.5 \times \left(\frac{\sharp chunks}{\sharp matched\_unigrams}\right)^3$$

$$\text{METEOR} = \frac{10PR}{R+9P} \times (1 - Penalty)$$

2.0.2 Literature Approaches

In literature there is a very long list of neural and non-neural approaches [Shi et al., 2018] [Allahyari et al., 2017]. Here, some of the main neural based

approaches are briefly presented.

One of the first neural encoder-decoder approaches to text summarization has been presented by [Nallapati et al., 2016]. They initially show that an off-the-shelf encoder-decoder framework, used for machine translation, already outperforms the previous ones for text summarization. Other than that, they augment input data by concatenating to classical word embeddings part-of-speech tags, named-entity tags and tf-idf statistics. Following the previous work of [Gu et al., 2016] , they also overcome to the problem of out of vocabulary words by equipping the model with a switching decoder/pointer architecture where the decoder has a "switch" that decides between using the generator or pointing to a word in the source text. This mechanism will be discussed in detail later. They also presented a new standard training and evaluation dataset called "CNN/Daily Mail" dataset, which will be presented in detail in chapter 3. In their paper they also use a novel attention method called *hierarchical attention*. The intuition behind hierarchical attention is that words in less important chunks of the input should be less attended. Therefore, with chunk-level attention distribution $\alpha^{chk,t}$ and word-level attention distribution by

$$\alpha_{ij} \text{ scale, } ,t = \frac{\alpha_i^{\text{chk, },t} \alpha_{ij}^{\text{wd, } t}}{\sum_{k,l} \alpha_k^{\text{chk },t} \alpha_{kl}^{\text{wd },t}}$$

So, each input gets divided into chunks. An attention is applied to these chunks. This re-scaled attention will then be used to calculate the context vector using

$$z_t^e = \sum_{i,j} \alpha_{ij}^{\text{scale},t} h_{ij}^{\text{wd}}$$

They evaluate their approach on the aforementioned CNN/Daily Mail dataset using the ROUGE metric.

Since representations of the source articles were still believed to be suboptimal, several works aimed to improving the encoding and decoding process.

The work from [Zhou et al., 2017] proposes selective encoding for text summarization. The idea introduces a selective gate network into the encoder for

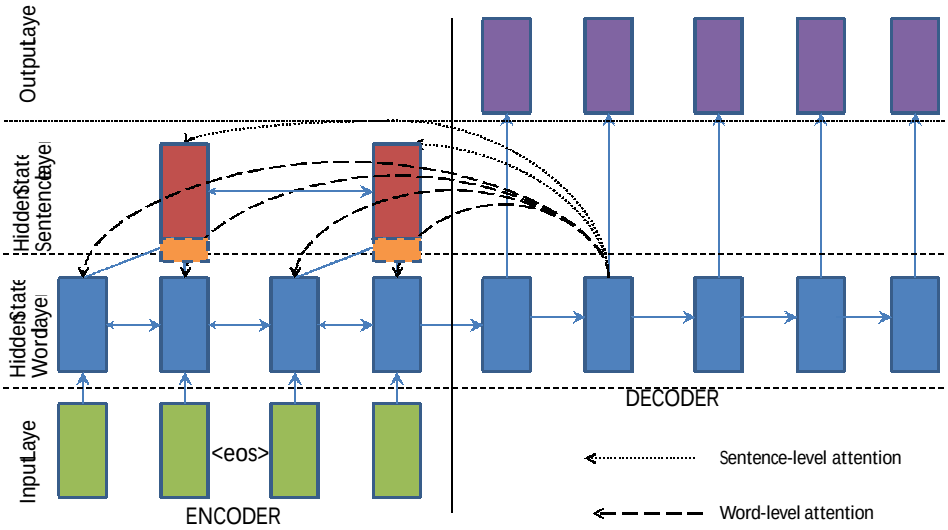

Figure 2.1: Hierarchical encoder with hierarchical attention. Attention weights at the word level are re-scaled by the corresponding sentence-level attention weights. Courtesy of [Nallapati et al., 2016]

the purpose of distilling salient information from source articles. A second layer representation called "distilled representation" is constructed by multiplying a selective gate to the hidden state of the first layer. The selective gate is a linear layers which applies transformations to a tensor being the sum of every hidden state of every step of the encoder. Such a gate network can control information flow from encoder to the decoder and can select salient information, boosting performance of the sentence summarization task (Figure 2.2).

Read-Again Encoding [Zeng et al., 2016] is an encoding mechanism driven by how human readers read an article. They often read it **several** times before writing a summary. This approach is motivated by the same idea behind

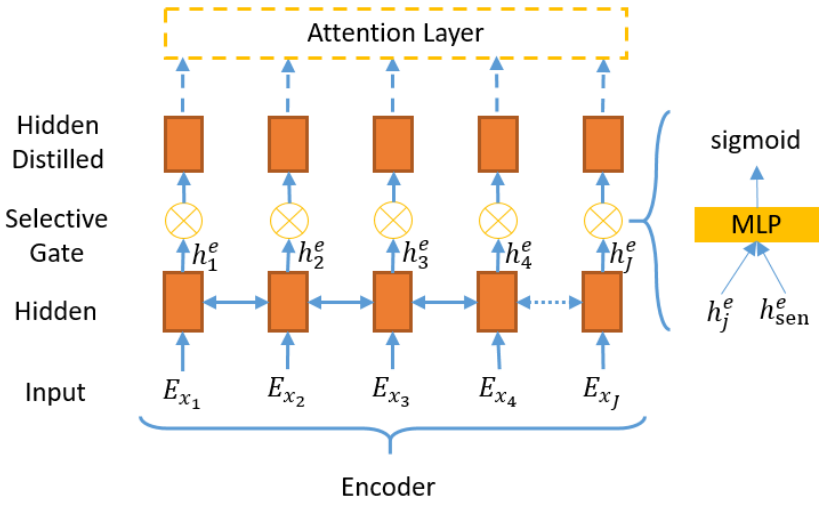

Figure 2.2: A visual illustration of selective encoding [Shi et al., 2018]

the attention mechanism, where the decoder network attends to the encoder state, thus re-reading the input source. The encoder reads the input two times. The first time an LSTM encodes tokens as $\left(h_1^{e,1}, h_2^{e,1}, \ldots, h_J^{e,1}\right)$ and $h_{\text{sen}}^{e,1} = h_J^{e,1}$, where $h_{\text{sen}}^{e,1}$ stands for a hidden representation compressing the sentence information. The second read, they use another LSTM to encode the source text based on the outputs of the first read. The encoder hidden state in the second read is updated by

$$h_j^{e,2} = \text{LSTM}\left(h_{j-1}^{e,2}, E_{x_j} \oplus h_j^{e,1} \oplus h_{\text{sen}}^{e,1}\right)$$

The hidden states of the second read are then passed to the decoder (Figure 2.3). Another approach proposed by [Paulus et al., 2017] aims to improve the decoder part of the model by sharing the embedding weights allowing the model to reuse the semantic and syntactic information in an embedding matrix during summary generation [Inan et al., 2016]. Having an embedding matrix named $\mathbf{W}_{emb}$ the matrix used in summary generation is:

$$\mathbf{W}_{\text{d2v}} = \tanh\left(\mathbf{W}_{\text{emb}}^{\top} \cdot \mathbf{W}_{\text{proj}}\right)$$

By sharing the embedding matrix and using the matrix of parameters W_{proj} the number of parameters is significantly less.

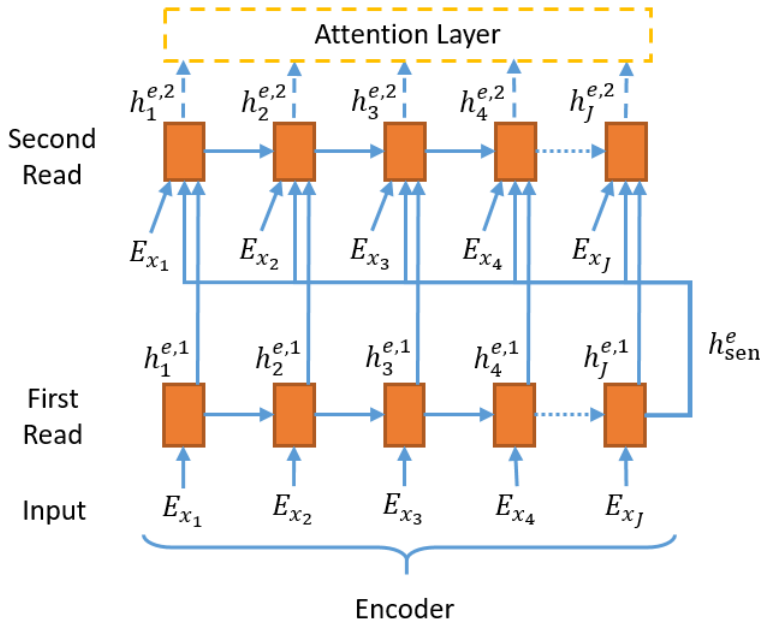

Figure 2.3: A visual illustration of read-again encoding [Shi et al., 2018]

Another original approach has been proposed in [Xia et al., 2017] where they try again to simulate how humans generate summaries. Human beings usually write first a draft and then they polish it based on the global context. Driven by this idea, authors propose an encoder-decoder framework where there are two decoders. The first decoder attends to encoder states and generates a draft. The second decoder attends to both the encoder and first decoder outputs. It generates the summary by exploiting information from *two* context vectors. This approach is called *deliberation network* and has also boosted the performance of seq2seq models in neural machine translation and abstractive text summarization.

A common problem of all encoder-decoder based approaches to text summarization is the *repetition problem*. Networks tend to have repetitions since the attention mechanism does not have a way to memorize the past attended part of the input and thus ignores the past alignment information [Tu et al., 2016] [Sankaran et al., 2016]. Models suffer of word-level and sentence-level repetitions, especially in datasets where summaries have sev-

eral sentences. In [Shi et al., 2018] and [Paulus et al., 2017] they propose *temporal attention* where they model the attention mechanism in order give lower attention to highly attended input tokens. Each attention distribution is divided by the sum of previous, dampening repeated attention.

Intra-decoder Attention [Paulus et al., 2017] in another technique to handle the repetition problem for long-sequence generations. here, the decoder not only attends tokens in the source text but also keeps track of the previously decoded tokens in a summary. The intra-decoder attention score is calculated as the normal attention score:

$$\alpha_{t\tau}^{d} = \frac{\exp\left(\mathrm{s}_{t\tau}^{d}\right)}{\sum_{k=1}^{t-1}\exp\left(\mathrm{s}_{tk}^{d}\right)}$$

where $\mathrm{s}_{t\tau}^{d}$ is the attention score given to each decoder hidden state. Then, the decoder-side context vector is computed by:

$$z_t^d = \sum_{\tau=1}^{t-1} \alpha_{t\tau}^{d} h_{\tau}^{d}$$

Another important approach is *coverage* [See et al., 2017] [Tu et al., 2016]. This simple yet powerful idea defines a coverage vector u_t^e as the sum of the attention distributions of the previous decoding steps, i.e. :

$$u_t^e = \sum_{j}^{t-1} \alpha_{tj}^{e}$$

It contains the accumulated attention information on each token in the source article during the previous decoding steps. The information of the coverage mechanism is then used in conjuction with the classical attention mechanism:

$$s_{tj}^{e} = (v_{\text{ align }})^{\top} \tanh\left(\mathbf{W}_{\text{ align }}\left(h_j^e \oplus h_t^d \oplus u_t^e\right) + \mathbf{b}_{\text{ align }}\right)$$

As a result, the attention of the current decoding time-step is aware of the attention during the previous decoding steps. Then this information is embedded into the loss function by adding *covloss*, which is a coverage-based loss value defined as:

$$_t = \sum_j \min\left(\alpha_{tj}^{e}, u_{tj}^{e}\right)$$

which forces the model to minimize the amount of attention given to already attended parts of the input.

Another approach is represented by [See et al., 2017], where they pro-

posed an hybrid method between abstractive and extractive models called *Pointer-Generator Network*. This type of model constitutes the core approach of this thesis to text summarization. It has been chosen because at the time of initial experiments it was the state-of-the-art model on text summarization. Their model facilitates copying words from the source text via *pointing* [Vinyals et al., 2015] improving handling of out-of-vocabulary (OOV) words. The basic model relies on the idea of encoder-decoder networks with a single-layer bidirectional LSTM with the classic attention mechanism [Bahdanau et al., 2015]:

$$e_i^t = v^T \tanh\left(\mathbf{W}_h h_i + \mathbf{W}_s s_t + \mathbf{b}_{\text{attn}}\right)$$
$$a^t = \text{softmax}\left(e^t\right)$$

with $\mathbf{v}, \mathbf{W}_h, \mathbf{W}_s$ and $\mathbf{b}_{attn}$ learnable parameters. The context vector is created as a simple weighted sum of the encoder hidden states h_t^*. The context vector is concatenated to the decoder state and passed through two classic linear layers in order to produce the vocabulary distribution P_{vocab}. So the probability of generating a word w is $P(w)$.
They use a custom loss function being the negative log likelihood of the *real* target word named w_t^* for the time-step t:

$$loss_t = -\log P(\mathbf{W}_t^*)$$

yielding the the overall loss for a single sequence as:

$$loss = \frac{1}{T} \sum_{t=0}^{T} loss_t$$

From this base model they extend the prediction mechanism by adding pointing mechanism, which allows the model both pointing words from the source text and generating words from its fixed vocabulary. The pointing mechanism is represented by the formula:

$$p_{\text{gen}} = \sigma\left(\mathbf{W}_{h^*}^T h_t^* + \mathbf{W}_s^T s_t + \mathbf{W}_x^T x_t + \mathbf{b}_{\text{ptr}}\right)$$

which takes the context vector, the decoder state and the decoder input at time t. Every input is multiplied by a different matrix of parameters learned during training. These transformations are summed up and passed through the sigmoid function yielding a probability value. This probability value

(p_{gen}) is interpreted as a soft switch between *generating* a word or copying a token from the input. An *extended vocabulary* is built by unifying the decoder vocabulary and the vocabulary created by only using words appearing the source text (so as to include also OOV words).
A new probability distribution is generated over the extended vocabulary:

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t$$

Each probability of words in the decoder vocabulary gets multiplied by the probability of generating it. Then, for every word in the source document, they sum the amount of attention received to the generation probability given from the decoder. If w is an OOV word, it gets $P_{vocab}(w) = 0$. This enables the usage of a small fixed vocabulary for the decoder; OOV words are pointed-copied from the input article.

In order to overcome to the repetition problem they employ the coverage mechanism already presented before. The maintain a *coverage vector* c^t keeping track of the sum of the attention given to each input token so far:

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$

The coverage vector is an unnormalized distribution and gets initialized with zeroes. Information about already attended tokens gets injected into the attention mechanism changing the equation to:

$$e_i^t = v^T \tanh\left(W_h h_i + \mathbf{W}_s s_t + \mathbf{W}_c c_i^t + \mathbf{b}_{\text{attn}}\right)$$

where w_c is a learnable parameter vector.
They also use the coverage loss which sums up the minimum between the attention vector and the coverage vector for each word. The new loss becomes:

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t)$$

A complete illustration of this approach is reported in Figure 2.4.

For all of their experiments they learn 128-dimensional word embeddings from scratch. They use a vocabulary of 50k words and 1-layer encoder and decoder based on LSTMs. They validate their approach on the CNN/Daily Mail dataset.

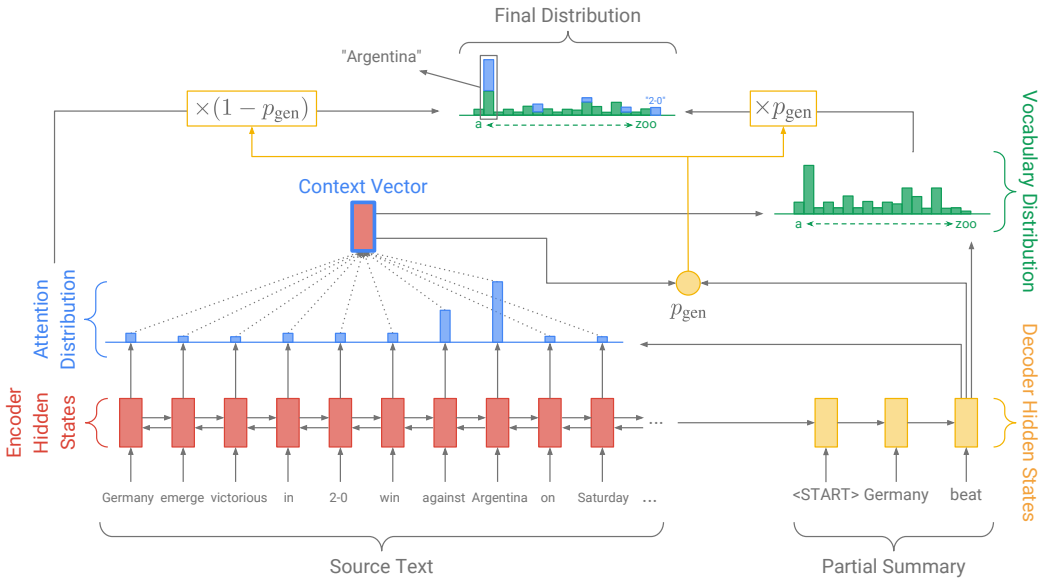

Figure 2.4: A visual illustration of the pointer-generator model. Courtesy of [See et al., 2017]. The encoder generates its hidden states (red). The decoder generates its hidden states (yellow). At each step the decoding process generates the attention distribution (blue), weights the hidden states and generates the context vector (red and blue). The decoder uses this information to generate the probability of generating a word from its vocabulary. Weights the generation distribution by the p_{gen} (green distribution) and sums the pointing generation after multiplying it by $1 - p_{gen}$, leading to the final distribution (on top of the image). This distribution is used by choosing the highest ranked word.

Another set of approaches uses reinforcement-learning to train models. An example is proposed by [Chen and Bansal, 2018] where they use two seq2seq models. The first is defined as an extractive model with the goal of extracting salient sentences from the input source. The second one is an abstractive model which paraphrases and compresses the extracted sentences into a short summary. They make use of convolutional neural networks to encode tokens and use an LSTM to encode and represent a document. They train the whole model by using standard policy gradient methods treating the two models as reinforcement learning agents. They evaluate their approach on the CNN/Daily Mail dataset.

[Paulus et al., 2017] present a new abstractive summarization model achieving state-of-the-art results on New York Times dataset. They introduce intra-temporal attention in the encoder and in the decoder, propose a new objective function by combining maximum-likelihood cross-entropy loss and rewards from policy gradient reinforcement learning to reduce exposure bias. Both models are trained optimizing directly the ROUGE score.

A parallel research direction goes beyond RNNs since they expose problems with vanishing gradients and computational costs in term of time and memory due how they handle long sequences. This research direction sheds light on convolutional neural network based encoder-decoder models. They can be parallelized during training and evaluation, the computational complexity is linear with respect to the length of sequences and this model can propagate gradient signals more efficiently than RNNs[Hochreiter et al., 2001]. A first approach is ByteNet [Kalchbrenner et al., 2016], where they adopts one-dimensional convolutions stacking on top of the hidden representation on the encoder CNN the decoder CNN.

Quasi-Recurrent Neural Networks (QRNN)[Bradbury et al., 2016] encoder-decoder based architecture is another approach which uses encoders and decoders made of convolutional layers and *dynamic average pooling layers* [Balduzzi and Ghifary, 2016] allowing computations to be completely parallel across batches and sequence time-steps. This framework has been demon-

strated to be effective and requiring less amount of time compared with LSTMs.
Several other approaches attempted to use convolutional neural networks for NLP and more specifically language modeling. One of the most famous models is proposed by [Vaswani et al., 2017] called the *transformer* which depends only on feed-forward networks and multi-head attention.

There are many other approaches, techniques and neural network architectures which will not be discussed here. For a more comprehensive overview refer to [Shi et al., 2018].

Chapter 3

Datasets for Text Summarization

Tackling the text summarization problem by means of deep neural networks, as every other machine learning approach, requires the availability of high-quality data for training and evaluation. Large datasets have driven rapid improvement in other natural language generations tasks, such as machine translation, where data size and diversity have proven critical for modeling the alignment between source and target texts. Access to large-scale high quality and annotated data, constitutes an essential prerequisite to make substantial improvement in summarization. In literature there have been published several datasets aimed for this specific task, most of them having a modest size. In this chapter every currently known dataset used to train and evaluate text summarization systems will be presented and analyzed.

Opinosis dataset [Ganesan et al., 2010]: it contains 51 articles, where each article describes a product and has 5 manually written gold summaries.

Large Scale Chinese Short Text Summarization Dataset (LCSTS): [Hu et al., 2015] this dataset is comprised of 2 million real Chinese short texts with short summaries and is built using the microblogging website SinaWeibo.

TCSum [Cao et al., 2015]: a dataset constructed by scraping and clustering tweets. It has only 1,114 documents.

scisumm-corpus [Jaidka et al., 2016]: published under the CL-SciSumm Shared Task for the CL-SciSumm 2018/2019. It has around 1,040 articles paired with some summaries. It has also a test set of 20 articles.

Live Blog Corpus for Summarization[Avinesh et al., 2018]: is a dataset for automatic live blog summarization posing new challenges for text summarization. It is comprised of about 180,00 documents.

TutorialBank [Fabbri et al., 2018]: an NLP general purpose set of datasets having 6,300 manually collected resources.

Legal Case Reports Data Set: it contains australian legal cases from the Federal Court of Australia (FCA).

TIPSTER: it has around 183 documents.

NEWS SUMMARY: a dataset with 4,515 examples scraped from the internet and uploaded on kaggle.

BBC News Summary: it contains 2,225 documents from the BBC news and it is used mainly for extractive text summarization.

sentence-compression [Filippova and Altun, 2013]: in the paper they describe an algorithm to collect data creating a large corpus of uncompressed and compressed sentences from new articles.

The Columbia Summarization Corpus (CSC) [Wang et al., 2011]: created with the newsblaster online news summarization system that crawls the web for news articles and clusters them on specific topics. It contains a total of 166,435 summaries with 2.5 million sentences over news in the 2003-2011 period.

Byte Cup Dataset: released by the Byte Cup 2018 International Machine Learning Contest for headline generation. It consists of 1.3 million pieces of articles with 1.1 million for training.

WikiHow-Dataset [Koupaee and Wang, 2018]: it contains crawled articles from the website wikihow.com. Each article is comprised of multiple paragraphs and each paragraph starts with a sentence summarizing it. By merging every sentence, one can create a summarization dataset with 200,000 pairs of articles and summaries.

Document Understanding Conference(DUC): [Harman and Over, 2004] it consists in newswire articles with human summaries. For each article there are multiple reference summaries.

Gigaword [Napoles et al., 2012]: it contains nearly 10 million documents from several newswire sources. It represents the largest dataset in this area. It does not contain summaries but news' headlines are often treated as summaries. Unfortunately this dataset is not free.

New York Times Corpus: it consists of news articles from The New York Times. Article and summaries make up a dataset of several hundred of thousand instances spanning from 1987 to 2007.

CNN/Daily Mail [Hermann et al., 2015]: this dataset is the most used and studied dataset in literature for text summarization. It includes CNN and Daily Mail articles. Each article comes with a bullet point list of phrases summarizing portion of each article. By concatenating the bullet points one can create single summary for each article. It is often used in its anonymized version where people names have been replaced with entity tokens. It is comprised of about 200,000 instances.

NEWSROOM [Grusky et al., 2018] : it is the largest known dataset created explicitly for summarization training and evaluation. It is comprised of nearly 1 million training instances, 100,000 validation

instances and 100,000 test examples. It has been created by crawling over 100 million pages from a set of 38 online publishers.

From the previous dataset list there are only few worth analyzing. They are NEWSROOM, CNN/Daily Mail, New York Times and DUC. Authors in [Grusky et al., 2018] make a very detailed analysis of these top summarization datasets which will be now presented. The results of this analysis strongly influenced the choice of which datasets to be used in this thesis.
Firstly, an important aspect of the NEWSROOM dataset is represented by its diversity. It contains summaries from different topic domains, written by many authors over the span of more than two decades. In order to analyze in a quantitative way these datasets some summary and article metrics are now presented.

Given an article text $A = \langle a_1, a_2, \ldots, a_n \rangle$ having tokens a_i and a corresponding article summary $S = \langle s_1, s_2, \cdots, s_m \rangle$, the set of fragments $\mathcal{F}(A, S)$ is the set of shared sequences of tokens in A and S.

Extractive Fragment Coverage

The first defined metric is the extractive fragment coverage which measures the extent to which a summary is derivative of its article.
$COVERAGE(A, S)$ measures the percentage of words in the summary which correspond to an extraction process from the source article:

$$\text{COVERAGE}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(A,S)} |f|$$

If a summary with 20 words has 10 words from its article text and has 10 new words, it will have $COVERAGE(A, S) = 0.5$.

Extractive Fragment Density

This metric measures how well the word sequence of a summary can be described as a series of extractions. It is similar to coverage:

$$\text{DENSITY}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(A,S)} |f|^2$$

Compression Ratio

It measures the simple compression ratio between the source article and the summary by dividing article's length to the summary length:

$$COMPRESSION(A, S) = |A|/|S|$$

The higher the compression the shorter the summary is; thus the system needs to be able to capture precisely the critical aspects of the article.

3.1 Analysis of Datasets Diversity

Extractive Fragment Density, Compression Ratio and Extractive Fragment Coverage have been used to profile and analyze the aforementioned summarization datasets. An extensive visual overview is reported in Figure 3.1 and Figure 3.2.

Figure 3.1 shows several types of datasets. Most diverse dataset sources are certainly *latimes.com* and *abcnews.go.com*. Specific ones are *telegraph.co.uk* and *cnn.com*. By merging all these types of sources NEWSROOM constitutes several techniques of summary generation leading to a higher quality dataset.

Figure 3.2 shows density and coverage across the top 4 datasets used for text summarization. The analysis shows how each dataset exhibits different human summarization strategies. Datasets are sorted by their median compression ratio. DUC's median compression ratio (47:1) is higher than every analyzed dataset. CNN/Daily Mail and New York Times have more extractive summaries. CNN/Daily Mail exhibits a high and specific coverage value, while New York Times has more diverse coverage behavior. NEWSROOM shows a very diverse way of summarizing text by having coverage values ranging from nearly 0.0 to 0.9. It also shows several degrees of density values meaning that it contains both extractive and abstractive summaries.

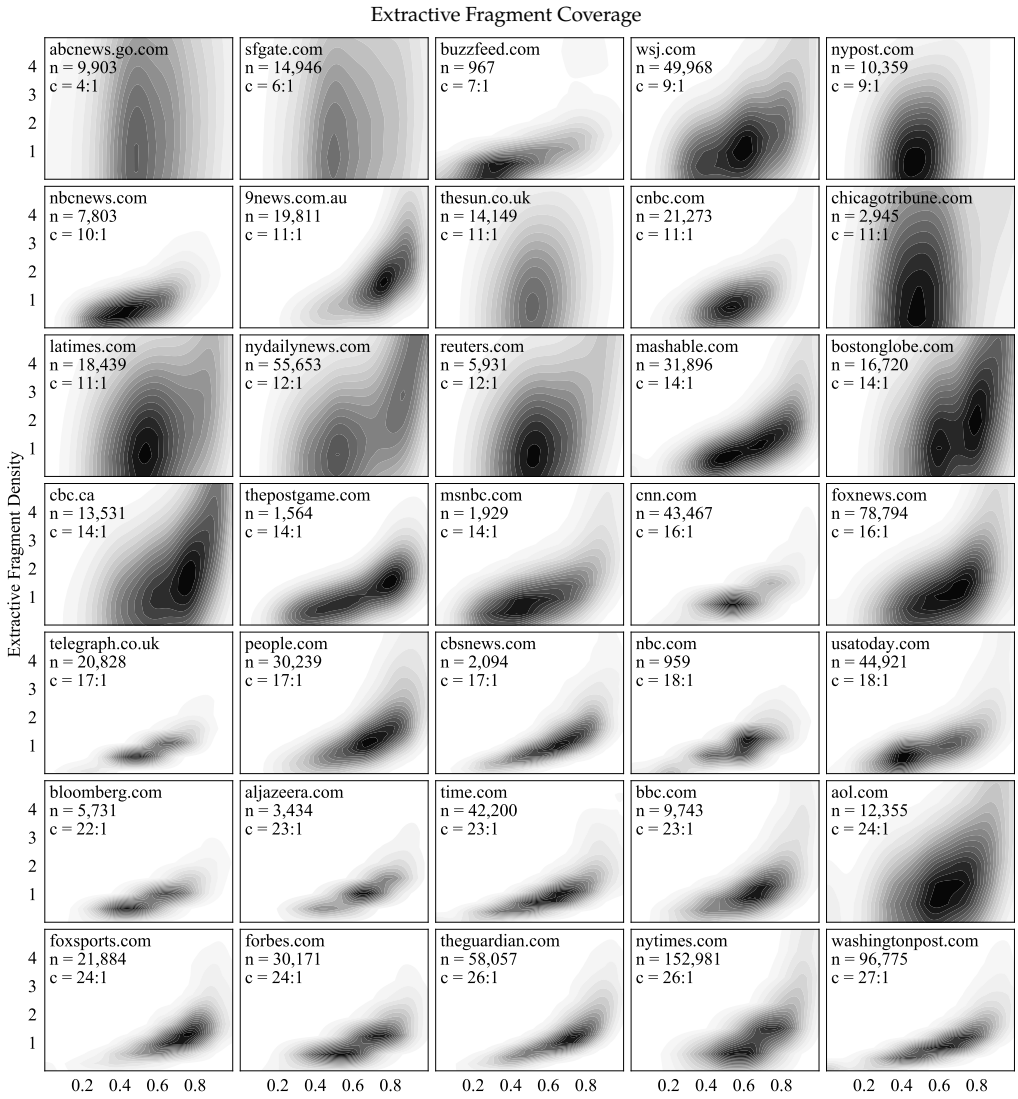

Figure 3.1: An overview of extractive fragment density and coverage across several only publishers used to create NEWSROOM. On the y-axis there is density and on x-axis the coverage. As shown from the plots NEWSROOM contains several types of summary generation techniques ranging from more extractive and coveraged ones such as *bostonglobe.com*, to less extractive such as *nbcnews.com*. On the top-left of each plot box n stands for the number of examples and c stands for COVERAGE. Plots are sorted by their median compression ratio. Washington post has the higher compression ratio of 27:1, while abcnews has the smallest one of 4:1. Courtesy of [Grusky et al., 2018].

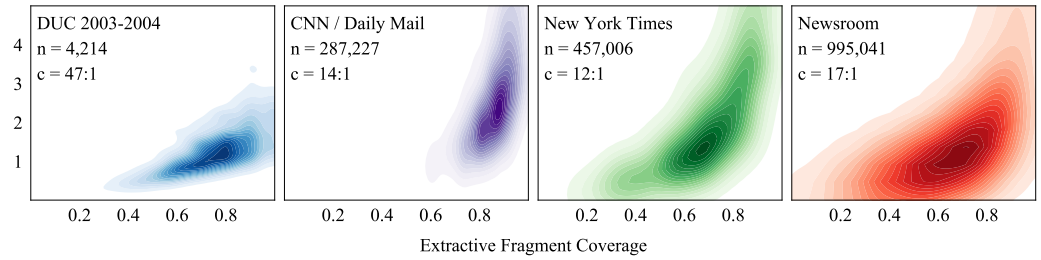

Figure 3.2: Density and coverage across the top 4 datasets used in text summarization. Courtesy of [Grusky et al., 2018].

Building systems capable of an abstractive way of reasoning and able to capture the meaning of an article text is the major challenge of automatic text summarization. Results from the previously reported analysis show that NEWSROOM is the most diverse, complete and large dataset currently published. Motivated by these results, NEWSROOM has been chosen as the main dataset for this thesis' research experiments. Moreover, being CNN/Daily Mail the standard dataset for automatic text summarization evaluation, it has been included in the experiments in order to compare this thesis' results with other results reported in literature.

The next chapter will present the proposed approach and all experiments' details.

Chapter 4

Experiments

In chapter 1 machine learning and deep learning have been presented. Their application to a specific area of machine learning called natural language processing has been extensively discussed in chapter 2. Chapter 3 listed every known text summarization dataset and discussed the main differences between them. In this chapter, the core approach of this thesis to text summarization is presented, alongside motivations behind every choice made. Experiments are then reported and discussed.

Pointer-Generator network architecture [See et al., 2017] proved to be a strong approach to text summarization. Its mechanism of pointing to source text tokens constitutes an original solution to the out-of-vocabulary problem, while the coverage loss is an effective loss term addition to overcome the repetition problem. One of the main cons of [See et al., 2017]'s approach is represented by the embedding method. This architecture uses 128-dimensional embeddings which are learned from scratch during training. Even though small-sized embeddings have been proven to yield very similar results compared to their higher dimension counterpart, a main motivation behind this thesis work is that low dimension embeddings *cannot* capture efficiently every semantic information of words. Learning high dimensional (1000+) embeddings from scratch during training would certainly slow down the overall process too much.

Driven by the excellent results brought by [Peters et al., 2018], a transfer learning approach is proposed. Instead of learning embeddings from scratch or even using fixed embeddings (GloVe, word2vec, ecc.), ELMo embeddings are fed into the pointer-generator network. Contextualized word embeddings should enable the model to pick up richer syntactic information and semantic information about words. This approach frees the pointer-generator model from the sub-task of learning how a natural language works and enables it to focus straightaway on the text summarization problem.

As the encoder reads the source text, a pre-trained ELMo model generates contextualized word embeddings. The encoder has two main sources to keep track of what has been read: its own memory and the inner information about past words injected into the current word contextualized embedding. This redundancy of information is helpful in order to generate richer encoder hidden states.

Each ELMo embedding is the result of a weighted sum of the ELMo's language model generated hidden states. In the experiments a task specific weighting of all bidirectional layers is performed by using learned from scratch weight parameters:

$$\mathbf{ELMo}_k^{task} = E\left(R_k; \Theta^{task}\right) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$

where $\mathbf{s}^{\text{task}}$ are softmax-normalized weights and γ^{task} allows the model to scale the entire ELMo vector.

Removing embedding learning from scratch allows the model to have several millions less parameters. On the other hand including ELMo adds 4 weights parameters which are tuned during training. The ELMo pre-trained model was trained on a dataset of 5.5 billion tokens consisting of Wikipedia and all of the monolingual news crawl data from WMT 2008-2012. The pointer-generator model is comprised of an encoder and a decoder using a single bidirectional layer LSTM cell with 256 as hidden size. The encoder gets 1024 dimensional embeddings from ELMo and it has a linear layer on top of the LSTM of size 512. Between the encoder and the decoder there is a tiny neural network called *Reduce State* with the aim of reducing the dimen-

sion of the tensor passed by the encoder to the decoder. It has two linear layers of 256 neurons with ReLU activation function. A layer compresses the cell state and the other the hidden state.
The decoder has a single bidirectional layer LSTM cell with size 256, followed by two linear layers of 256 neurons each. It also has an attention network implementing Bahdanau attention[Bahdanau et al., 2015] and including the coverage mechanism. Decoder's vocabulary size is set to 50,000 tokens. During training full teacher forcing is employed, by feeding into the decoder at timestep t the correct target words of the timestep $t-1$.

The original pointer-generator model has 21,499,600 parameters, while the current one has 19,349,593 parameters.

As loss function, the negative log-likelihood of the real target word is used:

$$loss_t = -\log P(W_t^*)$$

yielding the the overall loss for a single sequence as:

$$loss = \frac{1}{T}\sum_{t=0}^{T} loss_t$$

Batch size is set to 8 examples; a batch size of 16 has been tried and did not yield any particular improvement on the loss function while considerably increasing training time. Every example of the training set is cut at 400 tokens for encoding and 100 tokens for decoding. The decoder is ran at least for 35 steps. As the optimization algorithm, Adagrad is used with an initial learning rate of 0.15 and the Adagrad custom initial accumulator to 0.1.
Every LSTM cell is initialized sampling weights from the uniform distribution while keeping values in the range $[-0.02, 0.02]$. In order to prevent exploding gradients, gradient clipping has been applied with a max gradient 1-norm of 2.

As pre-processing step every article and summary is transformed to lower case and tokenized using *NLTK*[1]. Tags $<s>$ and $</s>$ are put around each summary in order to let the model be aware of the start and the end of each summary.

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

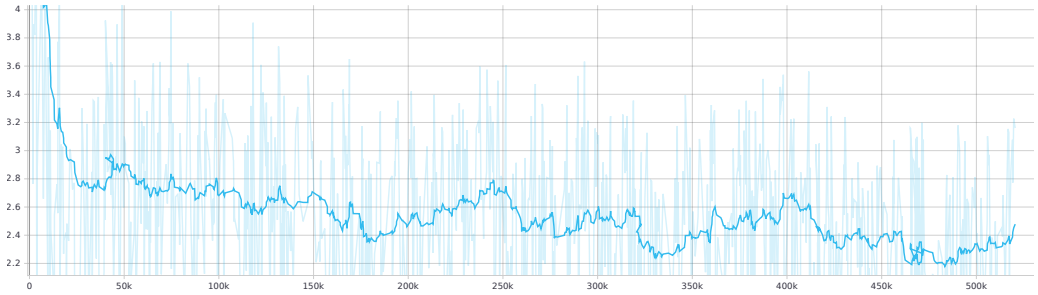

Figure 4.1: Training loss curve for the newsroom dataset. Raw curve (in the background) has been smoothed with a factor of 0.99 leading to the bold line.

At test time a beam search decoding is employed by using a beam size of 4.

4.1 Training results

This approach has been tested on two datasets: CNN/Daily Mail and Newsroom. Each dataset got the same text pre-processing and has been divided into several chunks of size 1,000 examples in order to better manage RAM and VRAM capacities. All experiments have been executed on a server with an NVIDIA TITAN Xp with 12GB of RAM, with CUDA version 10.0 and driver version of 415.27. A single training takes around 2,497 MB of memory.

As previously said CNN/Daily Mail has 287,226 training examples and 1,490 test examples, while Newsroom has 995,041 training examples and around 105,759 test samples.

A training of 455,000 iterations (12.6 epochs) has been performed on CNN/Daily Mail dataset, and a training of 520,000 iterations (4.1 epochs) has been performed on Newsroom. Training loss curve of newsroom and CNN/Daily Mail are reported in Figure 4.1 and Figure 4.2 respectively.

At first glance one can quickly notice the non-smoothness of both loss curves. Bold lines are the result of a smoothness with a factor of 0.99 of

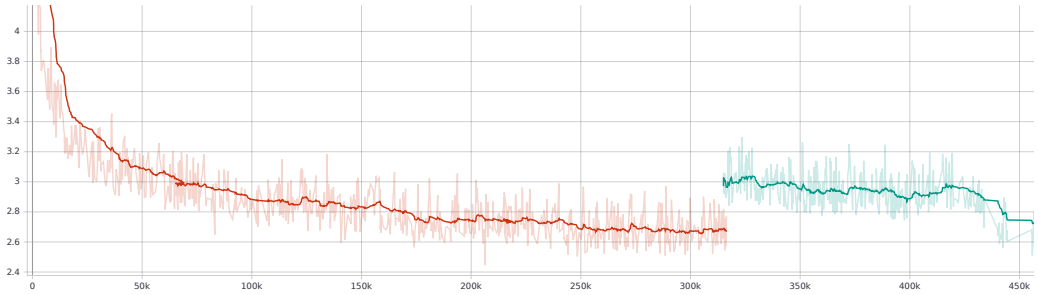

Figure 4.2: Training loss curve for the CNN/Daily Mail dataset. Raw curve (in the background) has been smoothed with a factor of 0.99 leading to the bold line. Red line shows training without coverage. Green line shows training with coverage loss enabled. Curve discontinuity (around 320k) is a normal behavior when using coverage loss.

the raw curves (showed in the background with semi-transparent lines). The first difference between the datasets is that Newsroom seems to cause more instable loss values with respect to CNN/Daily Mail, which can be interpreted as a more difficult loss surface landscape to harness with.

Another difference between the two training session is represented by the usage of the coverage loss. When training on Newsroom it has been noticed that the addition of the coverage loss in the complete loss function did not improve the model performance, but it even worsened the model's performance. Instead on CNN/Daily Mail the addition of the coverage loss (around 320k iterations) improved considerably the model's performance not so much with respect to the loss value but with respect to the ROUGE metric on the test set.

Figure 4.3 shows the comparison of the original pointer-generator model and the elmo augmented pointer-generator model with respect of their training loss values on CNN/Daily Mail. As one can clearly see the elmo augmented model is capable of converging much faster than the original model. This is because the initial model has to deal only with the summarization task and not with the language structure related rules (since the knowledge has already been "*injected*" by using elmo embeddings). Curiously the two

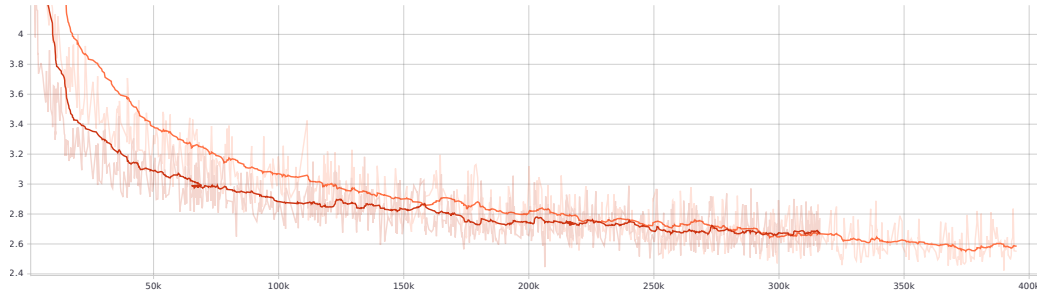

Figure 4.3: Comparison of two models. The orange one represents the original pointer-generator model, the red one represents the elmo augmented pointer-generator model.

	Learned weight
γ	0.35
LSTM-1	0.4140
LSTM-2	0.4690
LSTM-3	0.1169

Table 4.1: Softmax-normalized learned weights for the ELMo weighting equation. γ is the downstream task weight, while the other are per-layer weights.

models converge to the same loss value after $\sim$ 310 iterations.

The best performing model has been tested on the separate test set and its quantitative results are reported in table 4.2 in terms of ROUGE metric values. Quantitative results show the model is not as performing as the state-of-the-art models. In particular [Paulus et al., 2017] use a reinforcement learning based approach and [Liu, 2019] use extra training data other CNN/Daily Mail news.

Table 4.3 shows ROUGE metric values on the complete newsroom test set. The proposed model achieves better than state-of-the-art ROUGE-1 value for newsroom, and similar values for ROUGE-2 and ROUGE-L. [Grusky et al., 2018] applied the simple pointer-generator to newsroom achieving 26.04, 13.24 and 22.45 as ROUGE-1, ROUGE-2 and ROUGE-L respectively. In this work, the application of ELMo increases ROUGE-1, ROUGE-2

Paper	ROUGE-1	ROUGE-2	ROUGE-L
This work	38.96	16.25	34.32
[See et al., 2017]	39.53	17.28	36.38
[Paulus et al., 2017]	41.16	15.75	39.08
[Gehrmann et al., 2018]	41.22	18.68	38.64
[Liu, 2019]	**43.25**	**20.24**	**39.63**

Table 4.2: Rouge metrics on CNN/Daily Mail dataset. This work's results are reported in bold.

Paper	ROUGE-1	ROUGE-2	ROUGE-L
[Grusky et al., 2018] (Pointer-generator)	26.04	13.24	22.45
[Shi et al., 2018]	39.36	**27.86**	**36.35**
This work	**40.49**	27.15	34.11

Table 4.3: ROUGE metric values on the Newsroom test set. This work's results are reported in bold.

and ROUGE-L to 40.49, 27.15 and 34.11 respectively.

Table 4.1 shows softmax-normalized learned parameters for the γ weight of the ELMo equations and the three weights used to combine internal ELMo hidden states. As we can see the model gives roughly the same importance to the first two hidden states and a considerable less weight to the third layer activation. According to [Peters et al., 2018], the model favors syntactic information (captured at lower layers) instead of semantic information (captured at the highest layer). Finally, the learned specific downstream task weight γ is equal to 0.35.

In order to evaluate the model performance from a qualitative point of view a series of news, human generated summaries and software generated summaries are reported.

Article (tokenized) : -lrb- cnn -rrb- in case you have n't noticed , we 're in the midst of a medical marijuana revolution . given the amount of questions and mystery surrounding the science behind it , dr. sanjay gupta wanted to provide some insight . he 's been investigating medical marijuana for the last couple of years . his research has resulted in three cnn documentaries , culminating with " weed 3 : the marijuana revolution , " airing at 9 p.m. et/pt sunday . gupta opened up to questions on twitter . here 's what you wanted to know : . how does this affect me ? readers were curious about the effects of medical marijuana in easing symptoms of various ailments , asking how it could help with everything from life-threatening illnesses and neurological conditions to chronic pain and post-traumatic stress disorder . gupta pointed out the 10 diseases where medical marijuana research could have an impact and how it helped to stop one young girl 's severe seizures , and he also referenced other studies . he said that the potential benefits of medical marijuana for people suffering from ptsd is actually the subject of a federally approved study . the belief : it could suppress dream recall and allow those sufferers to focus on the present . there is also research into how the drug might affect the spasms associated with people who have multiple sclerosis . medical marijuana could also be an important option for those who rely on painkillers , as painkiller overdose is the greatest preventable death in the united states , gupta said . why do n't other doctors talk about it ? many of the questions around medical marijuana can be tied back to the fact that it 's just not discussed much . teaching about medical marijuana remains taboo in medical school . but gupta also presented the science directly , showing what your body on weed looks like . readers also wanted to know why people are so against the legalization of marijuana . decades of misinformation , gupta said . what are the drawbacks ? with medical marijuana so misunderstood , there naturally is a fear of potential side effects . there

are legitimate , long-term concerns with the developing brain , gupta said , and suggested that readers seek out the research of dr. staci gruber , who has conducted numerous studies on marijuana use and brain function . is n't it dangerous ? when confronted with questions rooted in fear about the dangers of medical marijuana , gupta confessed that he used to believe the same thing – that marijuana is dangerous without proven benefits . after extensive research , he said he changed his mind on weed .

Reference Summary:
cnn 's dr. sanjay gupta answers questions about medical marijuana . readers wanted to know how medical marijuana could ease symptoms of illnesses .

Generated:
dr. sanjay gupta has been investigating medical marijuana for the last couple of years . his research has resulted in three cnn documentaries , culminating with " weed 3 : the marijuana revolution " airing at 9 p.m. et sunday .

In this example the network uses an highly extractive approach replacing the personal pronoun "he" with "dr. sanjay".

Article (tokenized) : cairo -lrb- cnn -rrb- an egyptian court has sentenced 71 people to life in prison for their role in the august 2013 burning of a christian church in the giza province village of kafr hakim , state news reports . the virgin mary church was torched and looted by a mob , some of whom chanted against coptic christians and called for egypt to become an " islamic state , " one of at least 42 churches and many more businesses and homes targeted that august , the advocacy group human rights watch reports . others attacked included st. george church in sohag , a city south of cairo on the nile river , and prince tadros church in

fayoum , which is southwest of cairo , according to reports . in addition to those getting life sentences , two minors were sentenced to 10 years in prison and fined 10,000 egyptian pounds -lrb- about $ 1,300 -rrb- , egypt 's official egynews reported . most of those sentenced – 52 of the 73 defendants – were tried in absentia , with 21 already in prison , according to egynews . some blamed the church and other attacks on supporters of the muslim brotherhood , the islamist movement that backed mohamed morsy . morsy became egypt 's first democratically elected president following the ouster of longtime leader hosni mubarak . morsy was pushed out by egypt 's military . morsy was sentenced to 20 years in prison earlier this month after being convicted on charges related to violence outside the presidential palace in december 2012 . but he was acquitted of murder in the deaths of protesters . christian churches across egypt stormed , torched . journalist sarah sirgany reported from cairo and cnn 's greg botelho reported and wrote this story from atlanta .

Reference Summary:

2 minors were sentenced to 10 years in prison , in addition to adults getting life . 52 of the 73 defendants were sentenced in absentia . the virgin mary church was burned along with dozens of others in august 2013 .

Generated:

an egyptian court has sentenced 71 people to life in prison for their role in the august 2013 burning of a christian church . the virgin mary church was torched and looted by a mob , some of whom chanted against coptic christians and called for egypt to become an " islamic state " one of at least 42 churches and many more businesses and homes targeted that august , the advocacy group human rights watch reports .

In this example the network stays completely extractive providing more informations than the reference summary.

Article (tokenized) : ian bell rescued england from a nightmare start with a century in the first test against west indies that was pay back time for being dropped during his last caribbean series . after his classy 143 had steered england to 341 for five , after they had been 34 for three , bell said that thoughts of him losing his place in 2009 was a big motivating factor . bell said : ‘ it ’s been at the back of my mind for some time that i got dropped the last time i was here for a test series . so i wanted to make this one count and it was very satisfying to do so . i deserved to be dropped at the time and i did a lot of hard yards to get back in . ’ ian bell scored 143 on the first day of the first test against west indies in antigua on monday . the 33-year-old kisses the england badge on his helmet after his 22nd test match century . and bell rates his hundred one of the best of his 22 – only alastair cook -lrb- 25 -rrb- , out for another low score , and kevin pietersen -lrb- 23 -rrb- have scored more . bell shared century partnerships of 177 with joe root , who made 83 and another one of 130 with ben stokes who is 71 not out . ‘ it has to be right up there , ’ said bell . it does n’t get much better . the partnership with joe root was just what we needed . but the west indies bowled very well for the first two hours . and ben stokes made it a brilliant finish to the day . ’ bell was dropped from the side on his last visit to the caribbean in 2009 . bell put on 177 for the fourth wicket with joe root and added 130 for the fifth wicket with ben stokes . ’ it was my day today , but hopefully it will be someone else ’s day in the next innings . ‘ for me , it was nice - coming here in 2009 , this was where i got dropped after our shemozzle in jamaica , so i was determined to come out and play as well as i could . ‘ it was nice to come back and put things right . ’ bell raises his bat after being dismissed for 143 in the evening session on day one in antigua . kemar roach runs away to celebrate bowling alastair cook out for 11 on the first morning of the test . cook , who has n’t added to his record haul of 25 centuries since may 2013 , was bowled by kemar roach

for 11 falling again to a good length ball outside off stump . roach said : ‘ cook has a weakness there , everyone knows it . we let them get away a bit , but bell batted very well . ’ the other two england disappointments were a nervous-looking jonathan trott who lasted just three deliveries on his return from ‘ situational anxiety ’ , and gary ballance , who also fell to a poor shot .

Reference Summary:
ian bell hit 143 on the opening day of the first test in antigua . bell came in with england struggling , but took them to 341 for five . the 33-year-old was dropped during the 2009 series in the caribbean . bell says it was ‘ nice to come back and put things right ’

Generated:
ian bell scored 143 on the first day of the first test against west indies . ian bell was dropped from the side on his last visit to the caribbean in 2009 . ben stokes made it a brilliant finish to the day .

This example confirms an extractive approach wherethe network just replaces “bell” with “ian bell”.

Article (tokenized) : the former chairman of bradford city was linked to eight other fires before the valley parade blaze that killed 56 , a new book has claimed . author martin fletcher claims the devastating fire was not an accident and has revealed a sequence of other blazes at businesses owned by or associated with stafford heginbotham , the club ’s chairman at the time . after an inquiry , high court judge mr justice popplewell said the fire was not started deliberately and was caused by a discarded cigarette . west yorkshire police today said in a statement that they would review any fresh evidence surrounding the tragedy . scroll down for video . tragedy : the fire at bradford city ’s valley parade claimed 56 victims and injured 265 on may 11 , 1985 . then-bradford chairman

heginbotham -lrb- left -rrb- was linked to eight other fires before the valley parade blaze , it has been claimed . fletcher , a survivor of the blaze , spent 15 years researching the disaster in which his brother andrew , 11 , his father john , 34 , his uncle peter , 32 , and his grandfather eddie , 63 , were all killed . he was 12 at the time and after painstaking investigation of public documents has published his findings in his book : ‘ fifty-six – the story of the bradford fire ’ , serialised in the guardian . the fire broke out near half-time of bradford ’s game against lincoln on may 11 , 1985 , and was thought to have been started by a spark from a match or a cigarette dropped through a gap in the wooden frame of the main stand on to piles of rubbish which had been collecting for years . within four minutes the stand was completely ablaze . the fire was thought to have been started by a spark from a match or a cigarette dropped through a gap in the wooden frame of the main stand . the fire broke out near half-time in bradford ’s game . within four minutes the stand was completely ablaze . the fire , in 1985 , engulfed the stand within minutes . the 30th anniversary of the incident is approaching . then-bradford chairman stafford heginbotham -lrb- left -rrb- with mr justice popplewell in front of the burned stand . the judge said the fire was not started deliberately and was caused by a discarded cigarette . devastating : the deadly fire broke out near half-time during bradford ’s game against lincoln on may 11 , 1985 . a police officer uses his helmet to shield the heat from his face as he runs in front of the burning stand . his research includes claims that heginbotham , who died in 1995 at the age of 61 , had been connected to other fires at business premises over a period of the previous 18 years , which resulted in large insurance claims . the book , published on thursday , does not make any direct allegations but mr fletcher says heginbotham ’s history with fires , which he claims resulted in payouts totalling around # 27 million in today ’s terms , warranted further investigation . it was the worst stadium fire in the history of british football and prompted important safety changes

in grounds across the land . the final of game of the season should have been a day of celebration for bradford city . before kick-off the team were presented with the trophy for winning the third division title as 11,076 fans watched on . but instead , may 11 , 1985 ended in tragedy as flames engulfed the main stand at valley parade . the valley parade blaze is considered the worst stadium fire in the history of british football and prompted important safety changes in grounds across the land . the fire was noticed at 3.40 pm towards the end of the first half and within minutes the stand packed with 4,000 spectators was fully ablaze . the disaster left 56 supporters dead and a further 265 injured . in the years that followed , new legislation was introduced governing safety at the nation 's sports grounds . the tragedy brought about an unprecedented community spirit in the city , with , among many other initiatives , a fundraising drive for the burns unit at bradford royal infirmary . in 2010 , on the 25th anniversary , there were an estimated two thousand at the service in the city 's centenary square but this year there has been such a surge of interest that the club considered moving it to valley parade . ' could any man really be as unlucky as heginbotham ? ' wrote mr fletcher . ' from standing around with a bunch of kids and onlookers on a sunday afternoon in may 1967 , as his former foam-cushion business went up in flames , to standing on the pitch at valley parade 18 years later , making noises about smoke bombs while 56 people perished behind him . ' a 12-year-old martin fletcher is comforted by his mother susan in 1985 after the bradford fire disaster . he has spent years researching the disaster . survivor looks distraught with his burns visible -lrb- left -rrb- while another sits in front of a burned out stand . the release of the book coincides with the 30th anniversary of the disaster , with english football set for a minute 's silence ahead of all games on april 25 , which will be bradford 's closest home fixture to the date and is bound to cause much consternation in the west yorkshire city . 1 : may 1967 : fire in stafford heginbotham 's factory at cutler heights lane . 2 : april 1968 : fire at

genefoam ltd , managing director is heginbotham . 3 : august 1970 : store-room explosion at matgoods , founded by heginbotham . 4 : dec 1971 : tenant fire at castle mills , cleckheaton , owned by heginbotham . 5 : august 1977 : fire at yorkshire knitting mills , in heginbotham-owned douglas mills building . 6 : dec 1977 : fire at coronet marketing factory . coronet a subsidiary of tebro toys , owned by heginbotham . 7 : nov 1977 : fire with toxic fumes at his douglas mills factory . 8 : june 1981 : fire in a plastics factory at douglas mills . no-one sought accountability for the fire once the official popplewell inquiry had recorded its findings following a series of hearings held less than a month after the blaze . but mr fletcher started to investigate nine years after the disaster , scouring back-copies of the local paper , the telegraph and argus for information . sir olive popplewell told sky news today that he was not aware of the previous links to fires in 1985 . he said they were a ‘ remarkable coincidence ’ but that it did not alter his ruling that the tragedy was an ‘ accident exacerbated by negligence ’ . he added : ‘ if we had been aware of these fires at the time of course they would have demanded further investigation , but i am not sure what they would have revealed . ’ former sports minister gerry sutcliffe said today that the new allegations do not justify a new inquiry in to the disaster . mr sutcliffe , mp for bradford south and deputy leader of bradford city council at the time of the tragedy , says he knew heginbotham ‘ flew by the seat of his pants ’ in terms of the finances of the club but remains convinced by the conclusion of the inquiry by high court judge mr justice popplewell that the fire was an accident . prime minister margaret thatcher visits the scene of the disaster with her husband dennis in 1985 . tributes were paid to the 56 who were killed by the fire on the same day bradford should have been celebrating . the remains of the stand at bradford ’s ground after the disaster on saturday may 11 , 1985 at valley parade . his many businesses included toy company tebro toys . he once described football as ‘ the opera of the people ’ . after an initial spell as bradford chairman

, he came back in 1983 when he and another local businessman , jack tordoff , saved bradford from the receivers , buying the club for around # 30,000 each . heginbotham resigned as chairman in 1988 due to ill health . he died on april 21 1995 after a heart transplant . he was 61 . mr sutcliffe told press association sport : ‘ the inquiry by mr justice popplewell concluded that it was caused by a discarded cigarette in what was an old wooden stand and i have not heard anything to convince me that that was not the case . ‘ stafford heginbotham was one of those football club chairmen of which there were many at the time who flew by the seat of his pants . i was deputy leader of the city council at the time and he did fly by the seat of his pants when it came to paying the bill for the police and so on . ‘ but i think the inquiry was very thorough at the time and i do n’t think there needs to be another because of this . i do not believe there was any sort of cover-up and in fact the inquiry led to a lot of recommendations on stadiums that together with the taylor report came up with the right answers for football . mr fletcher started to investigate after a conversation with his mother nine years after the disaster , when she told him it had not been heginbotham ’s first fire . the charred remains of a bradford programme from the day of the fire in which 56 people were killed . ‘ there will always be speculation but i just think it was a tragedy that cost the lives of 56 people and injured many more , and has scarred the city for many years . ’ author martin fletcher writes about his own experience at the bradford disaster , and asks why chairman stafford heginbotham was not investigated : . ‘ why was it left to the 12-year-old who lost three generations and four members of his family , who was with over 40 people when they died within a 10-yard radius of him , who somehow got out of the stand , as the last person to get out , the only person to get out the front after being at the back , and the only person to survive the smoke ? ’ heginbotham ’s son , james , 47 , told the daily mirror today : ‘ when you actually do your homework and see what he did for bradford city football club it is a sickening

accusation . it is just absolutely ridiculous . ‘ he never recovered from the fire . the stress of it is what killed him eventually . it was a shock hearing this today , it was such a long time ago . ‘ he ’s no longer here to defend himself . it ’s a real shame it has come to this . ‘ he devoted his entire life to that club . he saved the club on two separate occasions . ’ i have never seen anybody more passionate about anything than he was about that club . ’ according to the independent , it would be up to home secretary theresa may to order a new inquiry but this would not be unlikely to happen before the election next month .

Reference Summary:
author martin fletcher claims the devastating blaze was not an accident . book reveals fires at other businesses owned by or associated with then club chairman stafford heginbotham . mr fletcher , a survivor of the blaze , says his findings warrant investigation . an inquiry into fire found it was an accident caused by discarded cigarette . west yorkshire police will consider any fresh evidence that comes to light .

Generated:
author martin fletcher claims the fire was not an accident . author said the fire was not started deliberately and was caused by a discarded cigarette . west yorkshire police said in a statement that they would review any fresh evidence surrounding the tragedy .

This example contains a false generated summary. The network mistakenly confuses the author Martin Fletcher with the high court judge by reporting a false citation by the author. This may be caused by the very high attention score given to the author throughout the article.

Article (tokenized) : it seems to happen after every single world cup that someone near to the top of english cricket gets the sack . whether it be duncan fletcher , david lloyd , whoever , someone always seems to pay

the price for our abject failure to compete on the global one-day stage . paul downton is just the latest victim . it is high time that we looked at the entire structure of english cricket and the reasons we are incapable of producing the types of cricketer required to succeed in international one-day cricket . paul downton has left his role as managing director of england and wales cricket board . moores -lrb- from left to right -rrb- , downton and england captain alastair cook during the press conference . the system is broken and it desperately needs fixing if england are ever going to compete in limited-overs cricket again . make no mistake , sacking paul downton does not mean some magic wand has been waved and all will be well . there are deep systemic problems . too many of the current selectors and coaches are from the old guard and have failed to keep pace with the dramatic changes that have occurred in one-day cricket in recent years . from downton to coach peter moores , to selectors james whitaker and angus fraser , these are players from a different era who have been painfully slow to react to situations . downton -lrb- left -rrb- and peter moores -lrb- right -rrb- pose during a press conference at lord 's cricket ground . downton leaves his role as ecb managing director following england 's shameful performance at world cup . even the sacking of alastair cook as one-day captain was a slow and protracted affair . they should have acted sooner . like so much of english sport , our cricket system seems afraid of the maverick player and that has to change . in my view there is still too much bad blood around for kevin pietersen to return , with the fall-out from the book causing untold damage to so many relationships that it just seems inconceivable for him to return . nasser hussain believes the sacking of cook as one-day captain was a slow and protracted affair . i hate it when things become about one individual and believe me , everything that has gone wrong with english cricket has not been down to the handling of kevin pietersen ; however badly that has been done . but what is does highlight is the continued fear of the maverick player . from david gower to kevin pietersen , english

cricket has shown distrust for anyone who does not conform or fit into a safe stereotype even though all the players who are thriving on the world stage are non-conformists who do things differently . you would n't find brendan mccullum or ab de villiers audacious shot - making in an mcc coaching manual but you will not find two better strikers of a cricket ball on the planet . english cricket needs to learn to embrace the maverick . pietersen has rejoined surrey in the hope of earning a place in the ashes squad for england this summer . hussain hopes whoever is appointed to replace downton is close to the modern game and ahead of the curve . i have been asked if i would like to take on downton 's job and the short answer is ' no ' . i will do anything i can to help the ecb and english cricket but there are far better candidates than me out there . whether it be michael vaughan , andrew strauss or anyone else , i just hope whoever they appoint is close to the modern game and ahead of the curve . we have been behind it for far too long.most importantly , it is time to wake up and look at the big picture . stop finding individuals to scapegoat and carry the can and start looking at the system that is failing to produce the kind of cricketers we need .

Reference Summary:

nasser hussain believes it 's time to look at the structure of english cricket . hussain believes paul downton is just the latest victim to take the flack . he feels the system is broken and desperately needs attention and fixing . alastair cook 's sacking shows we are painfully slow to react problems .

Generated:

paul downton has left his role as managing director of england and wales cricket board . it is high time that we looked at the entire structure of english cricket and the reasons we are incapable of producing the types of cricketer required to succeed in international one-day cricket .

This example shows that the network is capable of attending first to a

middle more important sentence and then to a previous less important sentence.

Article (tokenized) : the creator of the best-selling alex rider spy novels has accused fellow children 's author david walliams of dumbing-down fiction and failing to ‘ challenge ’ young readers . anthony horowitz said books by the comedian – who was the uk 's top-selling children 's author last year – are ‘ witty and entertaining ’ but nowhere near ambitious enough . the 59-year-old novelist and screenwriter singled out walliams ' gangsta granny for criticism , as well as the diary of a wimpy kid books by author jeff kinney . scroll down for video . author anthony horowitz -lrb- pictured right -rrb- said books by comedian david walliams -lrb- left -rrb- – who was the uk 's top-selling children 's author last year – are ‘ witty and entertaining ’ but nowhere near ambitious enough . he insisted that such writers should challenge their young readers and not be ‘ afraid of powerful stories or serious ideas ’ . he suggested they should follow the example of authors such as john green and ‘ write up for children , not down to them ’ . green 's book , the fault in our stars , tells the story of two teenagers who fall in love while they are both dying of cancer , and was the top-selling title of 2014 , with more than 870,000 copies sold . horowitz , who is also the principal writer on the itv period detective drama foyle 's war , singled out gangsta granny , with its breaking-wind jokes and a character who perpetuates the myth that children ‘ automatically dislike reading ’ . writing in the times educational supplement , he said he was troubled that children 's books ‘ seem to have come full circle ’ . walliams made over # 7million in 2014 from book sales including the boy in the dress and demon dentist . he added : ‘ to some extent , narrative fiction was reinvented by jk rowling – it 's hard to believe that children were n't challenged by books that stretched to 760 pages – and a phalanx of writers crested on her success : michael morpurgo , philip pullman , jacqueline wilson , malorie blackman , eoin colfer , darren shan

... and me . ‘ today it is jeff kinney and david walliams who top the bestseller lists , with books that are witty and entertaining but nowhere near as ambitious . ’ he added : ‘ speaking personally , i was a little sad when , in walliams ’ gangsta granny , we were told that among the old woman ’s vices , which included dribbling and farting : “ her house was stuffed full of books and she was always trying to get ben to read them , even though he loathed reading ” . ’ horowitz said ‘ the idea that children automatically dislike reading or that books belong to a more fusty , ancient generation is patently absurd ’ . his alex rider books , the first of which was published in 2000 , have been ‘ brightly re-jacketed ’ , and are ‘ as popular as ever ’ . they have sold 19million copies in 30 different languages , he pointed out . horowitz said : ‘ quite recently , i addressed an audience of largely unpublished writers in new york and i stand by the advice that i gave them . ‘ write up for children , not down to them . we do n’t need to be afraid of powerful stories or serious ideas . the audience is there . ’ walliams last year made more than # 7million from book sales . the 43-year-old britain ’s got talent judge had four books in last year ’s top 50 bestsellers list . his latest book , awful auntie , sold more than 500,000 copies last year . the success of his books , which also include ratburger and billionaire boy , has also been boosted by television adaptations , such as the boy in the dress , which starred walliams , meera syal and jennifer saunders , and was shown on the bbc at christmas .

Reference Summary:

author anthony horowitz has accused david walliams of dumbing down . claims walliams fails to challenge young readers with unambitious books . horowitz argues authors should not be afraid of ‘ powerful stories or ideas ’

Generated:

anthony horowitz said books by the comedian are ‘ witty and entertaining ’ the 59-year-old novelist singled out walliams ’ gangsta granny for criticism . he insisted that such writers should challenge their young readers and not be ‘ afraid of powerful stories or serious ideas ’

Here again the network shows a fairly extractive approach by skipping less important parts of the sentences.

Article (tokenized) : desperate chicago cubs fans were forced to pee into plastic cups as lines for the bathroom at wrigley field stretched through the concourse and out into the stands on opening day . the bathroom debacle caused people to wait more than one hour to get to one of only two working bathrooms and according to those who were there , some were left with no choice but to relieve themselves while standing in line . traumatized fans uploaded pictures to social media showing the intense queues and plastic glasses filled with urine , as the modernization of wrigley field reduced the famous stadium to a potential public health disaster . scroll down for video . busting : the line for the restroom can be seen snaking its way to the bathroom which is visible in the top left of the picture . almost there : disgruntled and disgusted fans uploaded pictures to social media revealing their trauma in getting to the bathroom . indeed , the cubs admitted as much on sunday as thousands of the 35,000 fans in attendance struggled to get to the bathroom . ‘ opening day at wrigley field has always brought challenges with wait times and tonight was particularly extreme , ’ cubs spokesman julian green said in a statement published after the cubs had lost to the st. louis cardinals 3-0 . stinking the field and the stands out , the cubs and wrigley field claimed the problem began when two bathrooms on the upper deck ‘ went down temporarily . ’ this came at a time when the stadium was ‘ already experiencing issues with long wait times . ’ social media was flooded with complaints . opening day : a general shot of wrigley field during batting practice prior to a game between the chicago

cubs and the st. louis cardinals on sunday night . ‘ i ’m all for renovating , but i ’ve now missed two innings trying to get into the bathroom , ’ said james hillebrand in a tweet . while cubs fan russ stubenruch said , ‘ the lines were obnoxiously long compared to what you would normally get . ’ one fan who identified herself as libby , told deadspin that the whole scene was a horror show . ‘ yes , fans were peeing in cups tonight . they were peeing everywhere . i heard there were only two working bathrooms - i know the lines were horrific and i also know that some were shut down early on because toilets just would n’t flush . ’ cubs spokesman julian green said that ‘ with 35,000 fans showing up in the ballpark tonight , we were simply not prepared to handle guests during peak periods . we have high standards for service and we missed the mark tonight . ‘ we want to apologize to our fans for the inconvenience tonight , ’ he added . ‘ moving forward we plan to supplement the existing restrooms with additional portable units and will continue to monitor wait times . ’

Reference Summary:

furious fans uploaded pictures to social media of huge lines for bathroom . some left with no choice but to use plastic cups to relieve themselves . wrigley field is undergoing huge renovation - its first overhaul in 80-years . two bathrooms flooded - forcing thousands to queue for over an hour .’

Generated:

chicago cubs fans were forced to pee into plastic cups as lines for the bathroom at wrigley field stretched through the concourse and out into the stands on opening day . the bathroom debacle caused people to wait more than one hour to get to one of only two working bathrooms and some were left with no choice but to relieve themselves while standing in line .

Article (tokenized) : the time has come for australia 's favourite rock star-turned-politician peter garrett to sell his charming victorian terrace in randwick in sydney 's affluent eastern suburbs . the retired midnight oil lead singer , former politician and passionate environmental activist has put his family 's sydney home on the market and is hoping the stunning terrace will be auctioned off for at least $ 1.05 million . the 62-year-old has lived at the thoughtfully restored home for almost five years with his wife dora and three daughters , emily , grace and may . as to be expected the 193 centimetre rockstar 's home boasts beautiful high ceilings . scroll down for video . the retired midnight oil lead singer , former politician and passionate environmental activist is listing his charming victorian terrace in randwick in sydney 's affluent eastern suburbs . the fresh , white kitchen opens up to a beautiful tiled courtyard at the rear of the home , bathed in sunlight and perfect for entertaining . the three bedroom terrace is set over two levels with a beautiful balcony off the master bedroom , looking out onto the street . the property has been exquisitely renovated and ' achieving a beautifully balanced blend of period charm and contemporary touches , ' according to belle property randwick . the garretts live just a short distance from 101 acres of greenery at centennial parklands and are just a short bus ride to iconic beaches such as bondi , maroubra and coogee . it offers the ultimate sydney lifestyle for any homeowner , whilst being less than 7km to the cbd . the property was bought by garrett for $ 932,500 in 2010 while he was the minister for environment protection , heritage and the arts as part of the gillard government . upon entering the home you are greeted by the living areas , masterfully designed with a clever use of space to include a living room , dining room , study nook and kitchen . the kitchen is crisp and clean with white , a light blue splashback and state-of-the-art kitchen appliances . from the kitchen , french doors open up the home . with the doors open , the dining room is extended out

into the garden - filling the house with light and making the home even more spacious . the family of five also own a property in mittagong in nsw 's south coast , ensuring they can enjoy the best of both city and country dwelling . the home is on a peaceful , historic randwick north street facade , charismatically built into the slope of gordon street . the three bedroom terrace is set over two levels with a beautiful balcony off the master bedroom , looking out onto the street . the winding staircase leads up from the lounge room to the bedrooms . the downstairs rooms have polished floorboards . upon entering the home you are greeted by the living areas , masterfully designed with a clever use of space to include a living room , dining room , study nook and kitchen . from the kitchen , french doors open up the home . with the doors open , the dining room is extended out into the garden - filling the house with light and making the home even more spacious . as to be expected from the former environmental minister , the garden is filled with greenery despite the city landscape and concrete surroundings . as to be expected from the former environmental minister , the garden is filled with greenery despite the city landscape . the winding staircase leads up from the lounge room to the bedrooms . the master bedroom has beautiful french doors which also open out onto a balcony . it is peaceful space to watch the day go by – enjoying a view of the stunning city skyline over the classic wrought iron balustrade . the property has been listed by belle property randwick and will go to auction on may 16th . the three bedrooms are all carpeted and all but one have built-in wardrobes . the stunning french doors open up to ensure light-filled , airy bedrooms . the stark white bathroom might have been a battle as peter garrett has three daughters ! he also at 193 centimetres tall , may have struggled to fit in the bathtub .

Reference Summary:

australia 's favourite rock star-turned-politician peter garrett has put his family 's sydney home on the market . he is hoping the stunning terrace

in randwick in sydney 's east will be auctioned off for at least $ 1.05 million . as to be expected the 193 centimetre rockstar 's home boasts beautiful high ceilings . he bought the property in 2010 whilst he was a federal minister for $ 932k . he also owns a mittagong property .

Generated:
retired midnight oil lead singer peter garrett to sell his charming victorian terrace in sydney 's affluent eastern suburbs . the 62-year-old has lived at the thoughtfully restored home for almost five years with his wife dora and three daughters , emily , grace and may . as to be expected the 193 centimetre rockstar 's home boasts beautiful high ceilings .

Here the network jumps to the middle of the article to gather some adjectives and then jumps to the middle of the first sentence in order to report the most important sentence of the article.

Article (tokenized) : a 19-year-old model in california has taken to bathing in pig 's blood in a desperate – and bizarre – attempt to maintain her youthful appearance – despite the fact that she is a vegetarian . in the latest episode of mtv 's true life , entitled ' i 'm obsessed with staying young ' , freelance model chanel details her unique skincare method , which she tries to justify by claiming that ' thousands of years ago people did this and it worked ' . ' my greatest fear is getting old and developing wrinkles and extra loose skin , ' chanel says . ' so to prevent that from happening i must do whatever it takes . ' scroll down for video . model of insecurity : 19-year-old chanel says she is terrified that the effects of aging will stop her modelling career in its tracks . ' i ca n't stop ' : chanel -lrb- l -rrb- picks up the blood from a local butcher . chanel insists that , despite her young age , she is already starting to show signs of aging , which she is convinced will soon put a stop to her modeling career altogether . her desperation to remain looking as young as possible for as long as possible has driven chanel to a number of extraordinary lengths

, most notably covering her body in animal blood , a process she insists will ‘ keep the skin looking soft and tight ’ . her family , skeptical of the benefits ask for proof , but chanel is unable to offer anything which backs up her bizarre claims . the model also shows off a myriad of other ‘ beauty fixes ’ she uses , including a daily shot of garlic powder , balsamic vinaigrette , sage , and olive oil - which she claims helps to ‘ preserve your body ’ - as well as a skincare regimen that involves washing her face eight to ten times per day . she also claims to have eaten placenta , gone on fasts and put her body through many detoxes in her quest to stay young . in the thick of it : chanel seems unperturbed by the process , dunking her hand into the container of blood and examining the contents before pouring it on to herself . bloody bizarre : the teenager picked up the blood at a local butcher who asked her if she would prefer pork or beef , to which she replies , ’ i do n’t eat meat ’ the teen ’s grandmother , lois , tries her best to dissuade chanel of her latest endeavor , saying she could be risking her well being . ‘ i ’m sorry my grandmother is worried about my health , ’ says chanel . ‘ but now that i ’m this close to my blood bath , i ca n’t stop . ’ after picking up the pig ’s blood from a butcher in a large bucket , chanel climbs into the tub wearing only her underwear and begins pouring the thick , red liquid all over her body . the model , who claims to be a vegetarian , describes how ‘ even the roughest parts of the body like the elbows feel soft , and i think i owe that to the blood ’ . later in the episode , chanel claims that though she believes the blood bath helped her self-esteem , she ‘ has no desire to do it again ’ . luckily for chanel , the experience of filming with mtv seems to have had a positive outcome . as she reveals in a follow up q&a with the channel , she was ‘ humbled ’ by the reactions she received . ’ i never believed i was that off my rocker , ’ she adds . ’ i think i owe that to the blood ’ : the teenager describes feeling her skin become softer as she soaks in the blood . other blood bathers : the infamous countess elisabeth bathory was said to have bathed in the blood of virgins to stay

young , and the benefits of its use are still championed today by the likes of kim kardashian . although the benefits are not proven , the concept of using blood as a beauty product is far from new . the idea of bathing in blood for youth may have originated with the story of elisabeth bathory , a hungarian countess who was alleged to have murdered hundreds of young women around the turn of the 16th century . folklore says that she bathed in the blood of virgins to maintain her youth . fast-forward to more recent - and less murderous - applications , kim kardashian 's ‘ vampire facial ’ went through a surge in popularity in 2013 after the reality star posted a photo of herself receiving the treatment , which involves drawing a person 's own blood , separating the platelets and then injecting it back into the patient 's face . a similar new beauty craze involved using a patient 's own blood to create a beauty cream , which was recently offered to patrons at the dorchester hotel in london .

Reference Summary:
chanel , from california , appeared on a recent episode of mtv 's true life . in the video she details her desperate attempts to avoid aging and prolong her modeling career for as long as possible . the teenager also eats placenta and washes her face ten times a day in order to keep her skin wrinkle-free .

Generated:
in the latest episode of mtv 's true life , entitled ‘ i 'm obsessed with staying young ’ , freelance model chanel details her unique skincare method . she tries to justify by claiming that ‘ thousands of years ago people did this and it worked ’

Article (tokenized) : one in five american women are willing to undergo dangerous beauty treatments in order to achieve the ideal look , despite the risks that these procedures pose to their health . according

to a new study , while just over half of women worry about the long term damage of beauty treatments , nearly a fifth would still pursue a treatment to get the right look - even it it proved hazardous to their health . seven per cent , meanwhile , have actually had allergic reactions . the survey , conducted by beauty research organization lqs and associates , looked at the lengths 1,000 american women go to in order to enhance their appearances or copy a celebrity , and the potentially disastrous consequences they might face in doing so , including hair loss , skin swelling , and overly painful procedures . the cost of beauty : women often do n't realize the dangers of salon treatments before sitting in the styling chair . respondents cited the hairstyles of actresses jennifer aniston and sandra bullock , as well as signer jennifer lopez , as celebrity favorites , with 53 per cent also noting that they get skin , beauty , and hair ideas from pictures and videos of famous personalities . but lqs and associates warns against attempts to emulate these looks on a regular basis . a-listers like beyonce , paris hilton , and selena gomez all sport weaves and extensions , but their beauty regimen is not always adaptable to the average person . ' many of those glamorous looks can do more harm than good , leading to long term problems , ' said laque gushon-harris , president of lqs and associates . ' unfortunately , most women are not even aware of the repercussions . ' mr gushon-harris explained that women who get weaves and extensions often deal with hair loss from traction alopecia , the gradual recession of the hairline due to tight hairstyles that pull at hair follicles over time . the survey also revealed that the top five beauty and cosmetic treatments favored by women in the us include manicures , pedicures , hair coloring , teeth whitening , and perms . however , frequently demonized treatments like tanning beds and facials also make the top 10 . good hair day : jennifer aniston -lrb- left -rrb- and sandra bullock -lrb- right -rrb- have the top two favorite celebrity hairstyles . long-term risk ? according to laque gushon-harris , president of lqs and associates , beyonce 's weave could be causing permanent damage to her

hair follicles . despite the dangers , the majority of women say beauty treatments make them feel better about themselves and improve their looks . nearly half believe these procedures boost their confidence . and in a country where one-sixth of women go to great lengths to look like their favorite celebrities , knowledge of dangers might not even stop them . when asked if they knew for certain that their treatments would cause permanent damage , four per cent said they would continue anyway and six per cent would continue , just not as much . most women , however , remain confident that their treatments are safe , and 67 per cent say they would stop if they learned what they were doing to themselves was unhealthy . it comes down to education , says mr gushon-harris , who travels the country teaching women about the dangers of cosmetic treatments . the study also found that 56 per cent of the respondents want more information on the side effects of their treatments and almost a quarter do n't have any idea of the dangers of their treatments . it 's a possible push for beauty salons to be more transparent . ' it is very important to be educated about whatever it is that you are doing to your body , ' he added . ' these damages can be lasting and life changing . '

Reference Summary:

american women look to celebrities for hair inspiration , often uneducated about the potential dangers of beauty procedures . many celebrities who wear weaves , such as beyonce , selena gomez and paris hilton , could be doing serious damage to their hair . jennifer aniston , sandra bullock and jennifer lopez were revealed as having the three most popular celebrity hairstyles .

Generated:

just over half of women worry about the long term damage of beauty treatments . nearly a fifth would still pursue a treatment to get the right look . respondents cited the hairstyles of jennifer aniston and sandra

bullock as celebrity favorites .

Here the network shows a fairly jumping approach which allows it to "compose" a single sentence from a set of chosen sentences.

Here are some examples from the newsroom test set:

Article (tokenized) : u.s. president barack obama on tuesday nominated former bank of hawaii chief executive allan landon to join the u.s. federal reserve ' s board of governors . lawmakers had been urging the white house to appoint someone with community bank experience to the board . if confirmed by the u.s. senate , landon would have a permanent vote on monetary policy and help implement the 2010 dodd-frank financial oversight law . landon , a partner with private investment fund community bancapital , served as chief executive officer of the bank of hawaii from 2004 until 2010. bancapital , based in portland , oregon , invests in the debt of community banks . according to the fund ' s website , landon , 66 , was responsible for " sourcing , evaluating and monitoring investments. " he worked as chief financial officer at bank of hawaii for four years before becoming ceo . previously , he had been the cfo at first american in tennessee and before that had worked at accounting firm ernst & young . the nomination comes as the fed prepares for its first interest rate increase since 2006 , which is expected around mid-year .

Reference Summary:
if confirmed , allan landon would have a permanent vote on monetary policy and help implement the 2010 dodd-frank financial oversight law .

Generated:
u.s. president barack obama nominated former bank of hawaii chief executive allan landon to join the u.s. federal reserve ' s board of governors . lawmakers had been urging the white house to appoint someone with community bank experience to the board . if confirmed by the u.s. sen-

ate , landon would have a permanent vote on monetary policy and help implement the 2010 dodd-frank financial oversight law .

Article (tokenized) : national institutes of health expert reviewers are finding flaws in the agency 's new study that connects heavy cell phone radiation to a slight increase in brain tumors in male rat ... a major u.s. government study on rats has found a link between cellphones and cancer , an explosive finding in the long-running debate about whether mobile phones cause health effec ... an australian judge on friday ruled in favor of a couple who refused to allow their 6-year-old son to undergo radiotherapy for a malignant brain tumor . a hawaii boy battling deadly brain cancer recently received news that he ' d receive an honor he ' s dreamt of during his short three years of life . hiv-infected cancer patients are much less likely to receive treatment for tumors than people who do n't have the virus , a large u.s. study suggests . a 7-year-old dallas boy with an inoperable brain tumor received a dream-room makeover thanks to a nonprofit and the generosity of his community . mandatory vaccination policies may encourage more healthcare workers to get annual flu shots and help prevent the spread of influenza to patients , a study in one texas health syste ... for elderly people in hong kong , long term exposure to fine-particle air pollution is tied to an increased risk of dying from many cancers , including breast , liver and pancreatic c ... neurosurgeons using lasers to treat brain cancer have discovered that the technique breaks down the blood-brain barrier , a finding that could lead to new treatment options for pati ... the risk of some childhood cancers might vary depending on where a child 's mother was born , a new study suggests .

Reference Summary:
brain cancer news articles and videos from foxnews.com 's health section.

Generated:
national institutes of health expert reviewers are finding flaws in the agency 's new study that connects heavy cell phone radiation to a slight increase in brain tumors in male rat ... a major u.s. government study on rats has found a link between cellphones and cancer .

This example shows a completely wrong ground truth summary. The network, instead, using again an extractive approach provides a better summary. There are several examples containing wrong reference summaries, which possibly hindered the learning process.

Article (tokenized) : much of the political world ' s focus is on the republican national convention , but in the background , there ' s big news on the horizon : hillary clinton will reportedly announce her running mate at an event in florida in just two days . agriculture secretary tom vilsack and sen. timothy m. kaine of virginia have emerged as the leading candidates on a longer list of finalists hillary clinton is considering for her vice-presidential running mate , according to interviews with multiple democrats with knowledge of her deliberations . although her list is not limited to those two , clinton has spoken highly of both in recent days to friends and advisers as she closes in on an announcement that could come as soon as friday . the post ' s article noted that clinton has sought advice on the matter from , among others , president obama . and while many try to read the tea leaves , clinton seemed to tilt her hand a bit on monday during an interview with charlie rose , which included the presumptive democratic nominee emphasizing " experience " as the key factor . " i am afflicted with the responsibility gene , " she added . the interview turned into a sort of word-association game . asked about kaine and his self-professed " boring " personality , clinton said , " and i love that about him . i mean , he ' s never lost an election . he was a world-class mayor , governor and senator , and is one of the most highly respected senators i know. " asked about hickenlooper , clinton said , "

first class. ” asked about warren , she added , “ amazing . i mean , what she has done in relatively few years to put the agenda of inequality front and center is something that i think we should all be grateful for . ”

Reference Summary:
the presumptive democratic nominee this week emphasized ” experience ” as the key factor . “ i am afflicted with the responsibility gene , ” she added .

Generated:
much of the political world ’ s focus is on the republican national convention , but in the background , there ’ s big news on the horizon : hillary clinton will reportedly announce her running mate at an event in florida in just two days .

Article (tokenized) : A healthy workforce is a more productive workforce, and finding ways to make taking care of everyday or acute care needs easier motivates many companies to offer on-site health care facilities. It2̆019s just one of many factors that helped elevate these nine companies to Fortune2̆019s list of the 100 Best Companies to Work For. Beyond enabling easier health care access for their employees, studies have found that on-site health care clinics can also lower medical costs for companies, encourage better long-term employee health, reduce absenteeism, and improve productivity, according to non-profit insurer The Alliance. On-site clinic capabilities vary by employer, and services range from last-minute appointments for a sinus infection to an annual physical by a primary care physician or even ongoing wellness options like a dietician or physical therapist. The service is also catching on among employers. Twenty-nine percent of companies with 5,000 or more workers offered an on-site or nearby health clinic in 2014, up from 24% in 2012, according to a national workplace study by consulting firm Mer-

cer. The survey also found that another 15% of large employers were weighing adding a clinic in the next two years. More than 40 companies on Fortune2̆018s list of the 100 Best Companies to Work For offer onsite health care centers, but many are only at their headquarters. Apart from the 11 hospital systems on our list, here are the nine companies that are currently offering this perk to employees at multiple locations.

Reference Summary:
They say it's cheaper for employees and better for business.

Generated:
a healthy workforce is a more productive workforce , and finding ways to make taking care of everyday or acute care needs easier motivates many companies to offer on-site health care facilities . it ' s just one of many factors that helped elevate these nine companies to fortune ' s list of the 100 best companies to work for .

This example shows a not properly human writtern summary. It would be better suited for a title than a summary.

Article (tokenized) : health and human services secretary kathleen sebelius should be " held accountable " for the botched rollout of the affordable care act ' s health insurance exchanges , rep. paul ryan , r-wis. , said today . asked by abc news on a conference call whether he agrees with virginia attorney general ken cuccinelli , a republican running for governor , that sebelius should step down , ryan said " i do agree with that. " " i ' d probably say that to her if i ever saw her in person , " ryan said . " it ' s becoming more and more clear by the day that they knew the rollout wouldn ' t go well. " since the exchanges opened for business on oct. 1 , the white house has acknowledged serious problems with the website that have made it difficult or impossible for some people to sign up for insurance . republicans have seized on the troubles with the

website as emblematic of larger problems with the law . and others , like cuccinelli , have called for sebelius to resign . despite the “ alarm bells , ” ryan said sebelius and the white house decided to proceed anyway . “ that looks like a call she made along with people in the white house , ” ryan said . “ and i do believe people should be held accountable. ” ryan has pressed sebelius for information about the website and the cost of the so-called “ tech surge ” that the obama administration planned to launch in order to fix it . in his latest letter , ryan writes , “ your continued silence on these important inquiries after refusing to testify raises serious questions about the administration ’ s commitment to transparency and accountability. ” two weeks before the nov. 5 virginia gubernatorial election , ryan joined cuccinelli on a conference call to denounce his democratic opponent terry mcauliffe ‘ s pledge to accept federal funds to expand medicaid in the state if he is elected governor . a recent nbc news/marist poll showed cuccinelli trailing mcauliffe by nine points in the race that is perhaps one of the most watched this year for its implications on presidential politics in 2016. cuccinelli , who opposes obamacare , said that accepting the federal funds is “ irresponsible budgeting. ” “ it ’ s really an obamacare expansion , it ’ s just in the medicaid program , ” cuccinelli said . “ medicaid has been one of the largest , most out-of-control parts of our state budget and we have to balance that budget every year , ” cuccinelli added . “ that ’ s why this decision on the front end is so important. ” ryan , who is the chairman of the house budget committee , is widely viewed as a rising figure in national politics , with an expertise in budget issues . though other prominent republicans such as sen. rand paul , r-ky. , plan to stump with cuccinelli in the last days of the campaign , ryan said today that his work leading a bipartisan group of senators and house members to come up with a budget compromise has tied up his schedule . “ i ’ m a little busy doing budget negotiations right now , ” ryan said . abc news ’ john parkinson contributed to this report .

Reference Summary:
(jewel samad/afp/getty images) health and human services secretary kathleen sebelius should be “ held accountable ” for the botched rollout of the affordable care act ’ s health insurance exchanges , rep. paul ryan , r-wis. , said today . asked by abc news on a conference call whether he agrees with virginia attorney general ken cuccinelli , a republican running for governor , that sebelius should step down , ryan said “ i do agree with that. ” “ i ’ d probably say that to her if i ever saw her in person , ” ryan said . “ it ’ s becoming more and more clear by the day that they knew the rollout wouldn ’ t go well. ” since the exchanges opened...

Generated:
health and human services secretary kathleen sebelius should be “ held accountable ” for the botched rollout of the affordable care act ’ s health insurance exchanges , rep. paul ryan , r-wis. , said today .

Article (tokenized) : Democratic Congressional candidate Krystal Ball Back when I used to obsessively cover Supreme Court justice Senate confirmation hearings for Above the Law, I used to think the dry hearings would be a lot more fun if nominees like Elena Kagan and Sonia Sotomayor had Facebook accounts to mine. This made me wonder what Senate hearings would be like decades from now when the prestigious nominees for government slots have social networking accounts that date back to their wild and crazy high school and college days. There2019s no need to wait decades, though. Krystal Ball, an ambitious candidate running for Congress in Virginia is 28, graduating from UVA in 2003. As a fellow 2003 college grad, I can attest to the rise in the use of digital cameras in those last couple years of college. Parties that involved alcohol and elaborate costumes were well-documented, with many of the photos being shared via digital photo sites like Flickr, Picasa, or, increasingly,

on Facebook. One of the photos that went viral Photos from a Christmas party Ball attended almost a decade ago resurfaced this month. A conservative blog posted the photos, featuring Ball in a sexy Santa costume, dragging her boyfriend, Rudolph the Red-Dildoed Reindeer, by a leash. It2019s unclear how the blog got its hands on the photos, whether through social-network-stalking Ball, or from a source. But they went viral 2014 Gawker, of course, has a nice slideshow. Ball herself was horrified at first, but realized she is on the leading edge of a generation that will have their photos and alcohol-fueled decisions stored and cataloged online. She recognizes that this is actually not solely a new-tech-based phenomenon 2014 Bill Clinton didn2019t need a Facebook account for his sex life to go viral 2014 but she advises her fellow young folk, and especially women, to be strong in the face of digitally-documented indiscretions going viral: Krystal Ball at a campaign event The tactic of making female politicians into whores is nothing new. In fact, it happened to Meg Whitman, one of the world2019s most accomplished business women, just last week. It2019s part of this whole idea that female sexuality and serious work are incompatible. But I realized that photos like the ones of me, and ones much racier, would end up coming into the public sphere when women of my generation run for office. And I knew that there could be no other answer to the question than this: Society has to accept that women of my generation have sexual lives that are going to leak into the public sphere. Sooner or later, this is a reality that has to be faced, or many young women in my generation will not be able to run for office. via Krystal Ball: The Next Glass Ceiling. This is really not limited to females. Plenty of my male friends have very embarrassing photos on Facebook too (despite my advice to them to groom their profiles). Ball is not accusing her detractors of violating her privacy, or apologizing for the photos. Instead she says that, moving forward, candidates and government nominees simply aren2019t going to be as flat to the world as they have been in the past. Thanks to digital

trails, everyone is going to have a far richer amount of data around them waiting to be explored. A few months back, IBM Chief Scientist Jeff Jonas, who made a business of mining data (and recently sat down with Forbes Video for an interview), suggested (hopefully) on the law blog Concurring Opinions that this kind of exposure will lead to a greater societal acceptance of our eccentricities: Unlike two decades ago, humans are now creating huge volumes of extraordinarily useful data as they self-annotate their relationships and yours, their photographs and yours, their thoughts and their thoughts about you ŏ2026 and moreŏ2026 How will mankind respond? Will people feel forced to modify their behavior towards normal only because they fear others may discover their intimate personal affairs? This is what Julie Cohen and Neil Richards have worried about ŏ2013 the ŏ201cchilling effect.ŏ201d Or, more optimistically, will the world become more tolerant of diversity? Will we be willing to be ourselves in a more transparent society? Personally, I shiver at the though of being on the hump ŏ2026 the hump of the bell curve. I hope for a highly tolerant society in the future. A place where it is widely known I am four or five standard deviations off center, and despite such deviance: my personal and professional relationships carry on, unaffected. via Concurring Opinions ŏ0bb Using Transparency As A Mask. Weŏ2019ll see how Ball fares in November, and will find out whether voters are accepting of her eccentric college day partying and whether her political career can carry on unaffected. Though, even if she loses, Iŏ2019m not sure we can deem that a rejection of a Gen Y candidate. It could just be the political make-up of her district. What I am sure of is that society will embrace Ballŏ2019s digital trail in a few weeks ŏ2014 Iŏ2019m willing to bet Iŏ2019ll see the ŏ201cKrystal Ball with raunchy reindeerŏ201d costume sported by a number of couples on New Yorkŏ2019s streets this Halloween. And Iŏ2019m sure there will be lots of photos taken to document the occasion. Hereŏ2019s Ball talking about the photos, via Dan Bigman:

Reference Summary:
virginia congressional candidate krystal ball 's racy photos offer a glimpse into the future of politics for the facebook generation .

Generated:
krystal ball , an ambitious candidate running for congress in virginia is 28 , graduating from uva in 2003. as a fellow 2003 college grad , i can attest to the rise in the use of digital cameras in those last couple years of college .

Article (tokenized) : Eleanor2019s efforts to fight racism , white supremacy and Jim Crow in the United States were more successful. Insisting that the country could not effectively champion democracy in the world when it practiced racial discrimination at home, she declared that because black Americans were 201cour largest minority, our attitude toward them will have to be faced first of all.201d Her efforts to fight lynching and end segregation in the military for men and women were backed up by her support for the N.A.A.C.P., her friendships with black activists and artists including Mary McLeod Bethune, Pauli Murray and Harry Belafonte, and her championing of Marian Anderson2019s concert at the Lincoln Memorial. Eleanor2019s prodigious activity served as her antidote to loneliness, anxiety and the periods of depression she called 201cGriselda moods.201d Cook weaves in a detailed account of her astonishing schedule; she spoke on the radio, gave lectures and talks and consulted speech specialists to help her lower and steady her voice. She traveled often, to meet people from migrant workers in California to wounded soldiers in Bora Bora. She kept up with art and music, and read 201cThe Grapes of Wrath201d and 201cFor Whom the Bell Tolls.201d Six days a week, from 1935 to 1962, she wrote a syndicated newspaper column called 201cMy Day,201d often composing it after midnight, in bed, in cars, on planes, or dictating to her secretary from the

bathtub. By 1939, these columns, which covered her travels, her home life, her views on human rights and her bracing, comforting words on courage in frightening times, reached more than four million readers. In December 1944, Cook suggests, Eleanor was the first American journalist to discuss Auschwitz. But in these tumultuous decades, which gave Eleanor the opportunity to play a major role on the national stage, her private life receded, submerged in the torrent of history. In 1940, Hickok wrote to her about the growing conflict between the real woman and the public image: ̆201cI . . . fought for years an anguished and losing fight against the development of the person into the personage,̆201d she lamented. ̆201cI still think the personage is an accident,̆201d Eleanor replied, ̆201cand I only like the part of life in which I am a person!̆201d Yet there are only a few places in the biography when she is introspective. In a 1943 letter to her old friend Esther Lape, she described the consequences of perpetually playing a role: ̆201cI find it hard to know sometimes whether I am being honest with myself. So much of life is play acting, it becomes too natural!̆201d On her life with Franklin, she concluded in the same letter, ̆201cthere is no fundamental love to draw on, just respect and affection. . . . On my part there is often a great weariness and a sense of futility in life but a lifelong discipline in a sense of obligation and a healthy interest in people keeps me going. I guess that is plenty to go on for onĕ2019s aging years!̆201d Cook shows Eleanor Roosevelt̆2019s final years as triumphant. After her husband̆2019s death in 1945, Truman appointed her to the United States delegation for the first United Nations assembly in London, where she helped write the Universal Declaration of Human Rights. Yet the last part of the biography emphasizes the personage rather than the person. No longer part of a loving community of women, Eleanor found her most intimate emotional relationships with younger men ̆2014 the writer Joseph Lash, and her personal physician, David Gurewitsch ̆2014 leading to complex triangular bonds with their wives that deserve deeper and fuller analysis.

Winding up Eleanor2019s majestic story in less than two pages, Cook doesn2019t describe her death from tuberculosis on Nov. 7, 1962. It was not an easy death; neither she nor her family was able to prevent painful, protracted and futile treatments her doctors hoped might save her. She had wanted a quiet private funeral, but got a spectacular state occasion. It2019s a tribute to Cook2019s rich portrait that after three enormous volumes, I still wanted to know more. The ironies of Eleanor2019s death make her life even more poignant and moving; their omission leaves her as a timeless and legendary figure, whose disembodied 201cglow,201d according to Adlai Stevenson2019s eulogy, 201ccontinues to warm the world.201d Eleanor Roosevelt was indeed a luminous beacon of courage and hope; yet the heroine of Cook2019s grand biography is not the remote icon, but the full-bodied, indomitable woman who welcomed life, as she put it, with 201can unquenchable spirit of adventure.201d Elaine Showalter is a professor emerita of English at Princeton. Her latest book is 201cThe Civil Wars of Julia Ward Howe.201d A version of this review appears in print on November 20, 2016, on page BR12 of the Sunday Book Review with the headline: First Lady to the World. Today's Paper—Subscribe

---

Reference Summary:
the long-awaited third volume of blanche wiesen cook ' s biography follows eleanor roosevelt ' s involvement with the united nations .

Generated:
eleanor ' s efforts to fight racism , white supremacy and jim crow in the united states were more successful . insisting that the country could not effectively champion democracy in the world when it practiced racial discrimination at home , she declared that because

This example shows an incomplete summary.

Article (tokenized) : PUYALLUP, Wash. ̆2013 Volunteers sifted through tons of paper at a recycling center Sunday, hoping to find items that Josh Powell may have dumped before killing himself and his two sons in a house fire last week. The search, involving about 20 people from Pierce County Search and Rescue, was expected to last all day.̈What happened last weekend really affected me. I could see smoke from the home from here,̈the recycling center's manager, Don Taylor, told KOMO News. ̈It would make me feel good if they found something here that could help answer questions about this whole terrible thing.̈Powell was the husband of missing Utah woman Susan Powell. He attacked his 5- and 7-year-old sons with a hatchet last Sunday, when a social worker brought the boys to his home for what was a supervised visit, and then ignited the house in a gas-fueled inferno. Some investigators have called the act an admission of guilt in his wife's death, but they're still looking for evidence on that front. They also located a storage locker where Powell had stashed a comforter that tested positive for the presence of blood in an initial test. Further lab tests are pending. Authorities received a tip that Josh Powell may have dumped some papers at the LRI recycling center in the days just before the deaths. The volunteers arrived Sunday morning and began searching through more than 20,000 pounds of recycled paper, junk mail and newspapers. A Pierce County sheriff's officer oversaw the effort. Susan Powell disappeared Dec. 6, 2009, while the Powells were still living in West Valley City, Utah. Josh Powell always claimed that on the night his wife disappeared he had taken his young sons on a midnight camping trip in subfreezing temperatures in the Utah desert, and that he knew nothing of her whereabouts. Less than a month after the disappearance, Powell moved the boys to his father's home in Puyallup, south of Seattle. Powell lost custody of his sons to his wife's parents late last year, after police said they had found child pornography on his father's computer. He later moved to the home outside Puyallup

where he killed his sons Feb. 5.

Reference Summary:
the pierce county sheriff 's office says a search has resumed at a recycling center near puyallup , wash. , for papers that josh powell is believed to have dumped before he killed himself and his two sons in a house fire last week .

Generated:
volunteers sifted through tons of paper at a recycling center sunday , hoping to find items that josh powell may have dumped before killing himself and his two sons in a house fire last week .

Article (tokenized) : Olympic leaders stopped short Sunday of imposing a complete ban on Russia from the Rio de Janeiro Games, assigning individual global sports federations the responsibility to decide which athletes should be cleared to compete. The decision, announced after a three-hour meeting via teleconference of the International Olympic Committee's executive board, came just 12 days before the Aug. 5 opening of the games. Ẅe had to balance the collective responsibility and the individual justice to which every human being and athlete is entitled to,ÏOC President Thomas Bach said. The IOC rejected calls from the World Anti-Doping Agency and dozens of other anti-doping bodies to exclude the entire Russian Olympic team following allegations of state-sponsored cheating. Russia's track and field athletes have already been banned by the IAAF, the sport's governing body, a decision that was upheld Thursday by the Court of Arbitration for Sport, and was accepted by the IOC again on Sunday. Calls for a complete ban on Russia intensified after Richard McLaren, a Canadian lawyer commissioned by WADA, issued a report Monday accusing Russia's sports ministry of overseeing a vast doping program of its Olympic athletes. McLaren's investigation,

based heavily on evidence from former Moscow doping lab director Grigory Rodchenkov, affirmed allegations of brazen manipulation of Russian urine samples at the 2014 Winter Games in Sochi, but also found that state-backed doping had involved 28 summer and winter sports from 2011 to 2015. But the IOC board decided against the ultimate sanction, in line with Bach's recent statements stressing the need to take individual justice into account. The IOC said the McLaren report had made no direct accusations against the Russian Olympic Committee äs an institution. Än athlete should not suffer and should not be sanctioned for a system in which he was not implicated,B̈ach told reporters on a conference call after Sunday's meeting. The IOC also said Russia is barred from entering for the Rio Games any athlete who has ever been sanctioned for doping. In a statement, the IOC said it would accept the entry of only those Russian athletes who meet certain conditions set out for the 28 international federations to apply. It also rejected the application by Russian whistleblower Yulia Stepanova, the 800-meter runner and former doper who helped expose the doping scandal in her homeland, to compete under a neutral flag at the games. However, the IOC added that it would invite her and her husband, Vitaly Stepanov, to attend the games.

Reference Summary:

olympic leaders have decided not to impose a total ban on russian athletes from the rio de janeiro games , which begin in less than two weeks .

Generated:

olympic leaders stopped short sunday of imposing a complete ban on russia from the rio de janeiro games , assigning individual global sports federations the responsibility to decide which athletes should be cleared to compete .

Article (tokenized) : some restaurant operators are scaling back expansion plans because of uncertainty about the expense of insuring employees under the new federal health-care law . the concerns are especially acute among smaller operators who are more likely to be on the cusp of the affordable care act 's requirements for increased coverage of workers . the doubt is adding to anxiety over other rising costs for items like ingredients at a time when diners are cutting back on eating out . sam ballas , chief executive of ecw enterprises inc. , owner of east coast wings & grill , a 26-unit chain in north carolina and texas , in ...

Reference Summary:
some restaurant operators , including white castle , are scaling back expansion plans because of uncertainty about the expense of insuring employees under the affordable care act .

Generated:
some restaurant operators are scaling back expansion plans because of uncertainty about the expense of insuring employees under the new federal health-care law . the concerns are especially acute among smaller operators who are more likely to be on the cusp of the affordable care act .

Article (tokenized) : aspen , colo. – authorities said saturday that a third suspect was arrested in the killing of a socialite in this colorado resort town , more than a week after a couple who rented the woman 's home were taken into custody . katherine m. carpenter , 56 , was apprehended friday night in the death of nancy pfister , 57 , the pitkin county sheriff 's office said . carpenter was being held without bond on suspicion of first-degree murder and conspiracy to commit first-degree murder . pfister was the daughter of the late betty and art pfister , longtime prominent aspen residents who co-founded the buttermilk ski

area west of town that 's hosted the winter x games multiple times . her body was found in an upstairs closet at her home on feb. 26 , but investigators have not said when or how she was killed . authorities said they arrested william f. styler iii , 65 , and nancy christine styler , 62 , at a lodge in basalt on march 3. they were staying at the lodge after apparently moving out of pfister 's home feb. 22 , the same day pfister returned from a vacation in australia , authorities said . the stylers , who are married , are expected to be formally charged monday with first-degree murder and conspiracy to commit first-degree murder . authorities said the couple rented pfister 's home during the fall . sheriff joe disalvo said his office has faced a few difficulties during the investigation . it 's the first homicide case in pitkin county in 12 years — not counting those classified as murder-suicides — and investigators have been conducting interviews in the courthouse , which is cramped . aspen is a ski resort town in the rocky mountains about 100 miles southwest of denver .

Reference Summary:
authorities said saturday that a third suspect was arrested in the killing of a socialite in this colorado resort town , more than a week after a couple who rented the woman 's home were taken into custody .

Generated:
authorities said saturday that a third suspect was arrested in the killing of a socialite in this colorado resort town , more than a week after a couple who rented the woman 's home were taken into custody .

An example of perfect match between the ground truth and the generated summary. It must be said that, in this case, the summary corresponds to the first two lines of the article and it is purely extractive.

Evaluating automatic summarization systems in an objective way is still a challenging problem because it blindly relies on the quality of the ground truth. As previously seen, system generated summaries often constitute a valuable and human readable alternative to the reference ones. Evaluating these type of systems qualitatively and quantitatively is important because given a corpus of text, even humans tend to generate fairly different solution summaries.

As already discovered by [See et al., 2017], the pointer-generator model tends to adopt an extractive approach most of the times. It must be said that training this type of systems on article news datasets causes the network to inherit the "human bias" of using the first and/or two sentences to represent the meaning of the whole article, and use the rest of the corpus to add secondary details. Following this human approach, the model too tends to copy the first one/two sentences of the article.

Massive datasets, such as Newsroom, have given and will continue to give a useful contribution to the research and development of systems capable of reading and comprehending short and long documents, with more or less information spread throughout the corpus. Even if these datasets represent very good reference summaries, several of such lack basic quality (as seen in the previous examples).

From the metric point of view, the ROUGE measure represents a too much quantitative way of measuring text summarization systems. In order to measure summaries in a qualitative manner, a new metric is needed. It has to deal with synonyms and whole phrases' meanings instead of unigrams, bigrams and L-grams as in ROUGE. A purely word overlapping approach can often guide research direction to new systems which are capable of maximizing metrics on test sets but not produce variegate and human like summaries.

Conclusions

The main goal of this thesis has been the application of pre-trained deep contextualized word embeddings to pointer-generator networks in order to train a machine learning system for automatic text summarization. In the first section machine learning and deep learning have been presented. Several recurrent neural networks as well as the sequence to sequence paradigm have been explored. The last part of the first section involved word representations and deep contextualized word embeddings from language model.

The second section introduced the text summarization problem by presenting both evaluation strategies and literature approaches.

The third section listed and analyzed several datasets used for text summarization.

The fourth section represents the core section of this work. It presents this work's approach as well as implementation details. Theoretical claims have undergone experimental analysis and results have been reported both in terms of quantitative metrics (with ROUGE values) and in a qualitative way (system generated summaries report).

Training has been conducted on two largest available datasets for text summarization: CNN/Daily Mail and Newsroom. Results shown that the ELMo enhanced pointer-generator model can increase substantially ROUGE metrics with respect to plain pointer-generator. On the newsroom dataset, the model achieved state-of-the-art ROUGE-1 value according to the test set. The model exhibited a more extractive way of generating summaries as already noticed by [See et al., 2017]. ELMo's weighting values have been

fine-tuned and showed the model's propensity to focus on word syntactic information rather than semantic one.
As future work it would be particularly interesting exploiting more novel architectures such as BERT [Devlin et al., 2018], the evolved Transformer [So et al., 2019], the Transformer [Vaswani et al., 2017] and OpenAI's GPT-2 [Radford et al., 2018] on huge datasets such as Newsroom.

The usage of the ROUGE measure is clearly not a perfect approach since it measures unigram, bi-gram and L-gram overlapping between system generated summary and the ground truth summary. Even with basic word preprocessing steps, the ROUGE metric lacks knowledge about synonyms and whole phrases meaning, causing a perfectly human understood system generated summary to have poor quality with respect to the ROUGE measure. More advanced measures are needed.

Bibliography

[Allahyari et al., 2017] Allahyari, M., Pouriyeh, S. A., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). Text summarization techniques: A brief survey. *CoRR*, abs/1707.02268.

[Avinesh et al., 2018] Avinesh, P., Peyrard, M., and Meyer, C. M. (2018). Live blog corpus for summarization. *CoRR*, abs/1802.09884.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

[Balduzzi and Ghifary, 2016] Balduzzi, D. and Ghifary, M. (2016). Strongly-typed recurrent neural networks. *CoRR*, abs/1602.02218.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (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.

[Bojanowski et al., 2016] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

[Bradbury et al., 2016] Bradbury, J., Merity, S., Xiong, C., and Socher, R. (2016). Quasi-recurrent neural networks. *CoRR*, abs/1611.01576.

[Cao et al., 2015] Cao, Z., Chen, C., Li, W., Li, S., Wei, F., and Zhou, M. (2015). Tgsum: Build tweet guided multi-document summarization dataset. *CoRR*, abs/1511.08417.

[Chan et al., 2015] Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. (2015). Listen, attend and spell. *CoRR*, abs/1508.01211.

[Chen et al., 2016] Chen, Q., Zhu, X., Ling, Z., Wei, S., and Jiang, H. (2016). Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR*, abs/1609.06038.

[Chen and Bansal, 2018] Chen, Y. and Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. *CoRR*, abs/1805.11080.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (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)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *MCSS*, 2:303–314.

[Daniluk et al., 2017] Daniluk, M., Rocktäschel, T., Welbl, J., and Riedel, S. (2017). Frustratingly short attention spans in neural language modeling. *CoRR*, abs/1702.04521.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[E. Rumelhart et al., 1986] E. Rumelhart, D., E. Hinton, G., and J. Williams, R. (1986). Learning representations by back propagating errors. *Nature*, 323:533–536.

[Elman, 1990] Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

[Fabbri et al., 2018] Fabbri, A. R., Li, I., Trairatvorakul, P., He, Y., Ting, W. T., Tung, R., Westerfield, C., and Radev, D. R. (2018). Tutorialbank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. *CoRR*, abs/1805.04617.

[Filippova and Altun, 2013] Filippova, K. and Altun, Y. (2013). Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

[Firth, 1957] Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. 1952-59:1–32.

[Freitag and Al-Onaizan, 2017] Freitag, M. and Al-Onaizan, Y. (2017). Beam search strategies for neural machine translation. *CoRR*, abs/1702.01806.

[Ganesan et al., 2010] Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Association for Computational Linguistics.

[Gehrmann et al., 2018] Gehrmann, S., Deng, Y., and Rush, A. M. (2018). Bottom-up abstractive summarization. *CoRR*, abs/1808.10792.

[Goodfellow et al., 2016a] Goodfellow, I., Bengio, Y., and Courville, A. (2016a). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Goodfellow et al., 2016b] Goodfellow, I. J., Bengio, Y., and Courville, A. (2016b). *Deep Learning*. MIT Press, Cambridge, MA, USA. `http://www.deeplearningbook.org`.

[Graves, 2012] Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385.

[Graves et al., 2013] Graves, A., Jaitly, N., and Mohamed, A. (2013). Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278.

[Graves et al., 2013] Graves, A., Mohamed, A., and Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778.

[Graves et al., 2014] Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *CoRR*, abs/1410.5401.

[Grusky et al., 2018] Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *CoRR*, abs/1804.11283.

[Gu et al., 2016] Gu, J., Lu, Z., Li, H., and Li, V. O. K. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *CoRR*, abs/1603.06393.

[Harman and Over, 2004] Harman, D. and Over, P. (2004). The effects of human variation in duc summarization evaluation. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 10–17, Barcelona, Spain. Association for Computational Linguistics.

[Harris, 1954] Harris, Z. S. (1954). Distributional structure. 10(2-3):146–162.

[He et al., 2017] He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

[Hermann et al., 2015] Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

[Hochreiter et al., 2001] Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

[Hornik, 1991] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

[Hu et al., 2015] Hu, B., Chen, Q., and Zhu, F. (2015). LCSTS: A large scale chinese short text summarization dataset. *CoRR*, abs/1506.05865.

[Inan et al., 2016] Inan, H., Khosravi, K., and Socher, R. (2016). Tying word vectors and word classifiers: A loss framework for language modeling. *CoRR*, abs/1611.01462.

[Jaidka et al., 2016] Jaidka, K., Chandrasekaran, M. K., Rustagi, S., and Kan, M.-Y. (2016). Overview of the cl-scisumm 2016 shared task. In *In Proceedings of Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries (BIRNDL 2016)*.

[Joos, 1950] Joos, M. (1950). *Description of Language Design*. The Journal of the Acoustical Society of America 22, 701.

[Józefowicz et al., 2016] Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

[Kalchbrenner et al., 2016] Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., and Kavukcuoglu, K. (2016). Neural machine translation in linear time. *CoRR*, abs/1610.10099.

[Koupaee and Wang, 2018] Koupaee, M. and Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

[Krizhevsky et al., 2017] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

[Lee et al., 2017] Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. *CoRR*, abs/1707.07045.

[Levy and Goldberg, 2014] Levy, O. and Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

[Lin, 2004] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[Lin et al., 2017] Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130.

[Liu et al., 2017] Liu, X., Shen, Y., Duh, K., and Gao, J. (2017). Stochastic answer networks for machine reading comprehension. *CoRR*, abs/1712.03556.

[Liu, 2019] Liu, Y. (2019). Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

[Luong et al., 2015] Luong, M., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

[Mani, 1999] Mani, I. (1999). *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

[McCann et al., 2017] McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. *CoRR*, abs/1708.00107.

[Mehler et al., 1988] Mehler, J., Pinker, S., and international journal of cognitive science, C. (1988). *Connections and symbols*. Cambridge, Mass. : MIT Press, 1st mit press ed edition. Reprinted from Cognition: international journal of cognitive science v.28, 1988.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

[Nallapati et al., 2016] Nallapati, R., Xiang, B., and Zhou, B. (2016). Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

[Napoles et al., 2012] Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 95–100, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Olah and Carter, 2016] Olah, C. and Carter, S. (2016). Attention and augmented recurrent neural networks. *Distill*.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Paulus et al., 2017] Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (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.

[Peters et al., 2017] Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *CoRR*, abs/1705.00108.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.

[Qian, 1999] Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Netw.*, 12(1):145–151.

[Radev et al., 2002] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. *Comput. Linguist.*, 28(4):399–408.

[Radford, 2018] Radford, A. (2018). Improving language understanding by generative pre-training.

[Radford et al., 2018] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018). Language models are unsupervised multitask learners.

[Ruder, 2016] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747.

[Rumelhart et al., 1986] Rumelhart, E, D., Mcclelland, J., and L, J. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations.*

[Salton, 1971] Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Sankaran et al., 2016] Sankaran, B., Mi, H., Al-Onaizan, Y., and Ittycheriah, A. (2016). Temporal attention model for neural machine translation. *CoRR*, abs/1608.02927.

[Schuster and Paliwal, 1997] Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681.

[See et al., 2017] See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

[Shi et al., 2018] Shi, T., Keneshloo, Y., Ramakrishnan, N., and Reddy, C. K. (2018). Neural abstractive text summarization with sequence-to-sequence models. *CoRR*, abs/1812.02303.

[So et al., 2019] So, D. R., Liang, C., and Le, Q. V. (2019). The evolved transformer. *CoRR*, abs/1901.11117.

[Srivastava et al., 2015] Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. *CoRR*, abs/1507.06228.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. NIPS*, Montreal, CA.

[Tu et al., 2016] Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016). Coverage-based neural machine translation. *CoRR*, abs/1601.04811.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Vinyals et al., 2015] Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc.

[Wang et al., 2011] Wang, W. Y., Thadani, K., and McKeown, K. (2011). Identifying event descriptions using co-training with online news summaries. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 282–291, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

[White, 1992] White, H. (1992). *Artificial Neural Networks: Approximation and Learning Theory*. Blackwell Publishers, Inc., Cambridge, MA, USA.

[Xia et al., 2017] Xia, Y., Tian, F., Wu, L., Lin, J., Qin, T., Yu, N., and Liu, T.-Y. (2017). Deliberation networks: Sequence generation beyond one-pass decoding. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1784–1794. Curran Associates, Inc.

[Zeng et al., 2016] Zeng, W., Luo, W., Fidler, S., and Urtasun, R. (2016). Efficient summarization with read-again and copy mechanism. *CoRR*, abs/1611.03382.

[Zhou et al., 2017] Zhou, Q., Yang, N., Wei, F., and Zhou, M. (2017). Selective encoding for abstractive sentence summarization. *CoRR*, abs/1704.07073.

Ringraziamenti

Vorrei ringraziare il mio relatore, il prof. Fabio Tamburini, per le critiche costruttive al progetto e i suggerimenti per la stesura della tesi. Vorrei ringraziare anche i miei genitori e il mio cane per il supporto fornitomi in questi anni di studio come studente fuori sede. Vorrei ringraziare Giulia, la mia ragazza, per essermi stata sempre accanto e avermi supportato anche nella fase finale di tesi. Infine, ringrazio naturalmente tutte le band musicali che hanno fatto da sottofondo allo sviluppo e stesura di questo elaborato.