Ukrainian Catholic University

Bachelor Thesis

Sentence Simplification in context of Automatic Question Generation

Author:
Dzvenymyra Yarish

Supervisor:
Prof. Jan Sedivy

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2019

Declaration of Authorship

I, Dzvenymyra YARISH, declare that this thesis titled, "Sentence Simplification in context of Automatic Question Generation" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ________________________________

Date: ________________________________

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Sentence Simplification in context of Automatic Question Generation

by Dzvenymyra YARISH

Abstract

Automatic question generation is a promising field of research. Sentence simplification is a key to the high-quality question generation. In this work we explore the existing sentence simplification approaches, present an valuable extension to the existing sentence simplification datasets and experiment with the latest seq2seq techniques. Our model makes use of transfer learning, which to our knowledge, is the first attempt to apply that method to sentence simplification. We evaluate our model against the rule - based approach and a baseline machine learning model and see improvements in the paraphrasing component of sentence simplification.

Acknowledgements

I am grateful to Prof. Jan Sedivy for inspiring and directing this thesis, Martin Matulík for sharing valuable insights on data mining and Czech Institute of Informatics, Robotics and Cybernetics for the opportunity to conduct a part of this research under their wing.

Also I would like to thank Ukrainian Catholic University and Faculty of Applied Sciences for cultivating such a rich academic environment, where me and this research could grow and thrive. In particular - huge thanks to Oles Dobosevych for always having answers to all of my questions.

Many thanks to Eleks for providing me with a scholarship to cover a part of my tuition fee.

Contents

List of Figures

List of Tables

List of Abbreviations

MLP	**M**ulti**L**ayer **P**erceptron
LSTM	**L**ong **S**hort -**T**erm **M**emory
GRU	**G**ated **R**ecurrent **U**nit
SoTA	**S**tate of **T**he **A**rt

Dedicated to my one and only bed, which never fails to support me.

Chapter 1

Introduction

1.1 Motivation

Question answering systems have gained a huge popularity over the last ten years. People interact with them on daily basis, often don't even noticing that interaction. The variety of question answering services is impressive, but developers usually don't pay much attention to the opposite action - question generation(QG). However, as the super advanced civilization in Douglas Adams' book (Adams, 1982) found out, the right question is as important as the correct answer. So not to get our ciris and google assistants throwing random numbers at us, we need to learn how to ask good, valuable questions. To free some time for focusing on fundamental questions we should delegate the task of asking routine, trivial ones to the machines. For example, when preparing a test for employees workplace security assessment, where people actually spend time on turning the narrative sentences in the security handbook into questions, or when any other factual knowledge evaluation is needed. Besides, knowing in advance all possible questions on a particular topic allows the apps, services, chatbots (and not very confident speakers) prepare answers beforehand and don't waste time on them while interacting with the user(audience). Especially in the field where every saved second of response time could grant a major edge on the competitors .

1.2 Problem framing

The main task is to generate all possible questions to an English narrative sentence, most probably encyclopedic or news extract. Questions should meet the following requirements:

- Grammatically correct
- Assessing content knowledge
- Not too vague
- Not too obvious
- The answer to the question must be found in the sentence

For example: *Francium was discovered by Marguerite Perey in France in 1939:*

- *Was Francium discovered by Marguerite Perey in France in 1939?*
- *What was discovered by Marguerite Perey in 1939?*
- *When was Francium discovered by Marguerite Perey?*

- *Where was Francium discovered by Marguerite Perey?*
- *Who was Francium discovered by?*

The most comprehensive attempt to complete this task was made by (Heilman, 2009). Their approach is composed of three stages: sentence simplification, question generation itself and question ranking. QG is a relatively trivial, well-described by rules task in linguistics. Th only issue there is the assignment of correct wh-words, which can be eliminated with the NER model. Fig.1.1 shows the algorithm for generating all possible questions to a given sentence used by (Heilman, 2009).

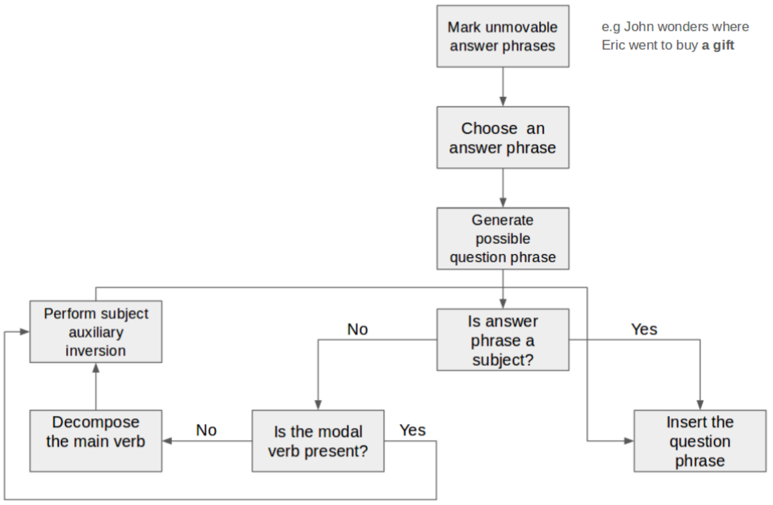

Mark unmovable answer phrases
e.g John wonders where Eric went to buy **a gift**
Choose an answer phrase
Generate possible question phrase
Perform subject auxiliary inversion
No
Is answer phrase a subject?
Yes
Decompose the main verb
No
Is the modal verb present?
Yes
Insert the question phrase

FIGURE 1.1: Question generation algorithm

It is boosted by succeeding ranking of generated questions by a logistic regression. Questions with low rank are then discarded.

The real challenge lies in the preprocessing step - sentence simplification. Table 1.1 clearly demonstrates why the adequate simplification of narrative sentences is crucial for easy-readable and comprehendible questions.

It is worth mentioning that sentence simplification alone is a worthy enterprise. It can facilitate information understanding by low-literacy people (children or non-native speakers) (Watanabe et al., 2009) as well as individuals with autism (Evans, Orasan, and Dornescu, 2014), aphasia (Carroll et al., 1999), or dyslexia (Rello et al., 2013). Besides, a simplification module could be used as a preprocessing step to enhance the performance of parsers (Chandrasekar, Doran, and Bangalore, 1996), summarizers (Klebanov, Knight, and Marcu, 2004), and semantic role labelers (Vickrey and Koller, 2008; Woodsend and Lapata, 2014).

1.3 Goals of the bachelor thesis

1. To explore the existing approaches to sentence simplification.
2. To enrich the existing sentence simplification datasets with new data.
3. To apply transfer learning to sentence simplification model.

TABLE 1.1: The importance of sentence simplification for QG: source [Wikipedia]

Original sentence	Simplified sentence
The Corporation of Leicester opposed the efforts of Charles I of England to disafforest the nearby Leicester Forest, believing them to be likely to throw many of its residents into poverty and need of relief.	Charles I wanted to destroy the Leicester Forest. The Corporation of Leicester opposed Charles I. It was likely to make many of its residents poor.
What did the Corporation of Leicester oppose to disafforest the nearby Leicester Forest ?	Who wanted to destroy the Leicester forest?
What opposed the efforts of Charles I of England to disafforest the nearby Leicester Forest ?	What Charles I wanted to destroy?
What did the Corporation of Leicester oppose the efforts of Charles I of England to disafforest ?	Who opposed Charles I?
What believed the efforts of Charles I of England to be likely to throw many of Leicester Forest residents into poverty and need of relief ?	What was likely to make many of Leicester Forest resident poor?
What did the Corporation of Leicester believe to be likely to throw many of Leicester Forest residents into poverty and need of relief ?	What was it likely to make many of Leicester Forest residents?

1.4 Thesis structure

The rest of the thesis is organized as follows: in Chapter 2 we explore previous approaches to sentence simplification, in Chapter 3 we provide all background information necessary to understand this work, in Chapter 4 we describe in detail our solution and Chapter 5 concludes the thesis with the comparisons of approaches and possible future work.

Chapter 2

Related works

2.1 General description

Sentence simplification has been a major field of interest for many computer linguists since 1990s. Being closely coupled with such problems as text summarization, reading comprehension and headline generation it introduces a thrilling challenge for those who decides to dive into it.

It is widely accepted that SS can be implemented by three major types of operations: splitting, deletion and paraphrasing (**Xu2016OptimizingSM**). The splitting operation decomposes a long sentence into a sequence of shorter sentences. Deletion removes less important or non-informative parts of a sentence. The paraphrasing operation includes reordering, lexical substitutions and syntactic transformations. An example of simplified sentence is show in Table 2.1:

TABLE 2.1: Example of sentence simplification

Original sentence	Simplified sentence
Due to recent heavy rainfall and subsequent flooding in Lviv, a city in the west of Ukraine, cars and buses were submerged in around a metre of water.	Lviv is a city in the west of Ukraine. It rained heavily there. The city got flooded. There was about one metre of water in the streets.

Despite the old age of the task, there is no benchmark developed for automatic evaluation of the quality of sentence simplification performed by various approaches. Thus, the person who tries to sort the solutions by their performance can only rely on the number of citations the paper describing the solution received or test the most promising ones on the hand-chosen set of sentences and human judge the quality of simplification with respect to the particular task - question generation in our case.

2.2 Rule-based approaches

Most of the conventional approaches focus on one of the mentioned earlier aspects of sentence simplification. Many solutions incorporate complicated rules for sentence splitting (Carroll et al., 1999; Chandrasekar, Doran, and Bangalore, 1996; Vickrey and Koller, 2008). Other achieve more simple sentences by substituting difficult words and complex phrases with the more common synonyms taken from WordNet or other corpus with synonymy information (Devlin, 1999; Inui et al., 2003; Kaji et al., 2002). More optimized definitions of rules and their application is used by (Cetto et al., 2018; Heilman, 2009). They isolate rules for dealing with different sentence complicating structures. For example, removing non - restrictive appositives, parentheticals and clause - level modifying phrases, extracting verb and noun modifiers, breaking coordinating and subcordinating conjunctions. The two systems similarly build constituency trees from original sentences using stanford tree parser (Chen and Manning, 2014) and then operate on them, cutting redundant parts and splitting into more simple sub trees. The first system makes use of a convoluted set of rules implemented as if-then-else, second uses a special tree regular expressions language (Levy and Andrew, 2006). The apparent drawbacks of such approaches, besides operating very poorly on the semantic component of simplification, is debugging complexity and difficult scalability.

2.3 Statistical approaches

Recent approaches view the simplification problem more holistically (Zhang and Lapata, 2017), aiming to excel at all parts of simplification simultaneously. In this case the task can be viewed as the monolingual text-to-text generation problem which closely resembles the statistical machine translation. (Zhu, Bernhard, and Gurevych, 2010) were the first who tried to apply the syntax-based translation model from (Yamada and Knight, 2001) to sentence simplification, which additionally performs simplification-specific rewrite operations. Another data-driven solutions include (Woodsend and Lapata, 2014) model based on quasi-synchronous grammar (Smith and Eisner, 2006), which allows to capture structural mismatches and complex rewrite operations; (Wubben, Bosch, and Krahmer, 2012) phrase-based machine translation (PBMT) model(Koehn, Och, and Marcu, 2003), which K outputs are scored according to the similarity measure to the original sentence; (**Xu2016OptimizingSM**) syntax-based machine translation model, which is trained on a large-scale paraphrase dataset PPDB (Ganitkevitch, Durme, and Callison-Burch, 2013) making use of simplification-specific objective functions.

2.4 Deep learning

After the tremendous success of neural networks in machine translation task (Cho et al., 2014b) it seemed logical to try to apply them to sentence simplification as well. A seq2seq encoder - decoder neural network achieved good results and established a healthy tradeoff between simplicity and meaning preservation (Nisioi et al., 2017a). The subsequent works proposed various adjustments and improvements to the base model. Two worth noting here are reinforcement component (Zhang and Lapata, 2017), where the model learns to maximize the reward function which depends on how closely the model outputs meet simplification requirements, and Neural Semantic Encoder architecture (Vu et al., 2018).

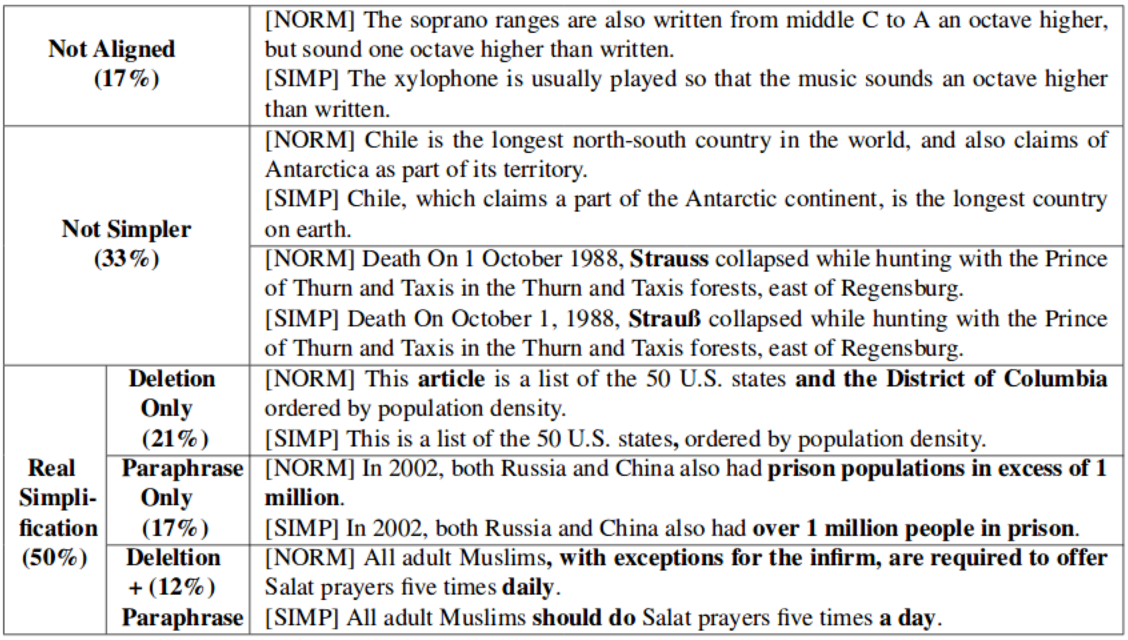
2.5 Available datasets

The recognised dataset for sentence simplification is Parallel Wikipedia Corpus (Coster and Kauchak, 2011) which contains 200K sentence pairs from English Wikipedia and Simple English Wikipedia. The pairs were created by automatic alignment of corresponding articles. However, as other researchers claim (Xu, Callison-Burch, and Napoles, 2015) the corpus possesses many weaknesses:

- Only 50% of sentence pairs are really aligned.
- Not sufficient degree of simplification in most cases (as shown by manual examination of random sample).

- Enormous vocabulary size (due to encyclopedic nature).

Those weaknesses can be clearly seen in Fig.2.1.



Not Aligned (17%)		[NORM] The soprano ranges are also written from middle C to A an octave higher, but sound one octave higher than written. [SIMP] The xylophone is usually played so that the music sounds an octave higher than written.
Not Simpler (33%)		[NORM] Chile is the longest north-south country in the world, and also claims of Antarctica as part of its territory. [SIMP] Chile, which claims a part of the Antarctic continent, is the longest country on earth.
		[NORM] Death On 1 October 1988, **Strauss** collapsed while hunting with the Prince of Thurn and Taxis in the Thurn and Taxis forests, east of Regensburg. [SIMP] Death On October 1, 1988, **Strauß** collapsed while hunting with the Prince of Thurn and Taxis in the Thurn and Taxis forests, east of Regensburg.
Real Simplification (50%)	**Deletion Only (21%)**	[NORM] This **article** is a list of the 50 U.S. states **and the District of Columbia** ordered by population density. [SIMP] This is a list of the 50 U.S. states**,** ordered by population density.
	Paraphrase Only (17%)	[NORM] In 2002, both Russia and China also had **prison populations in excess of 1 million**. [SIMP] In 2002, both Russia and China also had **over 1 million people in prison**.
	Deleltion + (12%) Paraphrase	[NORM] All adult Muslims**, with exceptions for the infirm, are required to offer** Salat prayers five times **daily**. [SIMP] All adult Muslims **should do** Salat prayers five times **a day**.

FIGURE 2.1: Aligned pairs of sentences from Parallel Wikipedia Corpus. [(Xu, Callison-Burch, and Napoles, 2015)]

As the result, models trained on that corpus, regardless of how advanced the architecture is, fail to perform good simplification. For quite a long time there was no substantial progress in sentence simplification with deep learning due to the poor quality of Parallel Wikipedia Corpus, until the new dataset, Newsela corpus (Xu, Callison-Burch, and Napoles, 2015), was released. It consists of 150K pair of sentences, extracted from the news articles(see). The researchers behind it claim that each original article was rewritten 3-5 times by the language experts for the readers with different levels of English. Therefore, the higher simplification quality is achieved, while the alignment of corresponding sentences from different complexity articles is still automatic.

For the sake of completeness, another SS corpus should be mentioned - (Xu et al., 2016). It was developed to use with SARI metric, but due to its modest size (2359 sentence pairs), it can be used only for tuning and testing.

TABLE 2.2: Newsela example sentence pairs

Normal sentence	Simplified sentence
American women will soon be free to fight on the front lines of battle and they will go with the public ′s support .	American women will soon be able to fight in wars .
Former Governor Jon Corzine , a Democrat , tried to make an issue of Christie ′s weight in the 2009 gubernatorial election.	Former New Jersey Governor Jon Corzine ran against Christie in 2009 .
She pointed out that black soldiers used to be forbidden from serving alongside whites and that Japanese-Americans were once segregated into their own battalions .	She pointed out that it used to be forbidden for black soldiers to serve alongside whites .
The results were as spectacular as they were nauseating .	The results were sickening .
As the economy begins to perk up and businesses start to hire , a lack of basic knowledge about mathematics could present a problem to people looking for work .	They are looking to hire new employees.

Chapter 3

Background information and theory

3.1 Neural Networks

Artificial neural network is an attempt to mimic the way human brain works. It is build of neurons, which are budndled into layers. One neuron has connections to some(or all - in that case it is a fully connected layer) neurons in the previous layer which resembles dendrites, with a weight coefficient assigned to each connection which controls how much information to take from it, imitating brain synapses. Each neuron outputs a signal, transformed by so-called activation function which adds non-linearity to the output and was developed to model the frequency of action potentials, or firing, of biological neurons. The simplest (vanilla) neural network architecture is called multilayer perceptron(MLP). It consists of three layers: input, hidden and output (Fig.3.1).

The growing interest around such net architecture is justified not only by the biological structure, but also by a solid math proof called the Universal Approximation Theorem ("Approximation with artificial neural networks"), which claims that every continuous function defined on a compact set of the nth-dimensional vector space over the real numbers can be arbitrary well approximated by a feed-forward artificial neural network with one hidden layer (with finite number of artificial neurons). This is a ground breaking statement, meaning that if one can accept that most classes of problems can be represented as functions, a neural network can, in theory, find a solution to all of them. In practice, of course, it is constrained by infeasible computations and ineffective optimization algorithms.

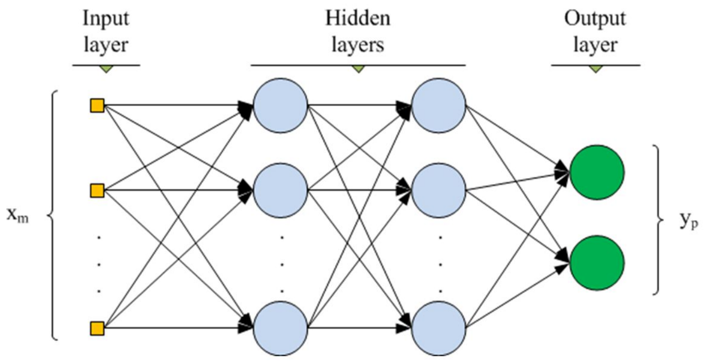


FIGURE 3.1: Schematic MLP architecture [Image source]

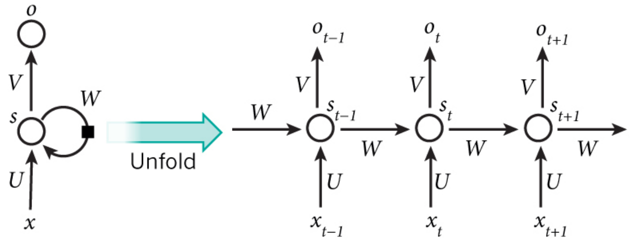


FIGURE 3.2: RNN architecture [Image source]

Still, neural networks now perform tasks which are almost impossible for conventional algorithms (object detection, speech recognition, text generation). Largely because neural networks are capable of learning most meaningful for a specific task data representations by finding hidden patterns and complex dependencies in high-dimensional data.

The training of neural networks is done in two steps: first the training data is forwarded through the network and the error between network predictions and ground-truth values is calculated, in the second step this error is backpropagated through the network and the weights are updated with respect to the error gradients.

The research in the field of neural networks and deep learning is currently progressing in two directions: creating a task-specific changes to the basic architecture to ensure the most beneficial handling of data and developing effective optimization methods to achieve the optimal convergence for SoTA results and reduce the computational complexity of training.

3.1.1 Recurrent Neural Networks

A feed-forward neural network defines a mapping $y = f(x;\theta)$ and learns the values of parameters θ which leads to the best function approximation (Goodfellow, Bengio, and Courville, 2016). In feed-forward neural networks information flows only in one direction: from input to output layer. That way the output of a neuron doesn't affect it, meaning there are no loops in the network, when the outputs of the model are fed into itself. However, in case of sequential data(time-series or text) $x^{(1)}, x^{(2)}..., x^{(T)}$ where $x^{(t)}$ is dependent on $x^{(1)}.,.., x^{(t-1)}$ a kind of memory is needed to efficiently process the sequence.

Recurrent neural networks ("Learning representations by back-propagating errors") introduce loops in the information stream, maintaining a certain form of memory, called hidden state h and exploit the idea of sharing the parameters between the units of the network. Equation below defines the value of the hidden state of the RNN:

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

The power of RNNs lies in mapping the arbitrary length sequence $x^{(1)}, x^{(2)}..., x^{(t)}$ to a fixed length hidden state vector $h^{(t)}$, so the model learns to include in it a summary of task-specific aspects of the previous elements in the sequence (Goodfellow, Bengio, and Courville, 2016). One typical RNN architecture is used for predicting a sequence of o values from seq of x values (Fig.3.2).

The following equations are applied for the forward pass through the model at every timestep:

$$a^{(t)} = b + Wh(t-1) \ + Ux^{(t)},$$

$$h^{(t)} = tanh(a^{(t)},$$

$$o^{(t)} = c + Vh^{(t)},$$

$$\hat{y} = softmax(o^{(t)})$$

where the trainable parameters are W and V weight matrices and b and c are bias terms. Here we use hyperbolic tangent function $tanh$ as the activation function. Softmax operation on the output converts it to the normalized probabilities of each possible value of the discrete variable, which depending on task can represent a word or a character. The log likelihood loss is computed between the target Y sequence and predicted $\hat{Y}$ and backpropagated through the network using back-propagation through time (BBBT) algorithm.

3.1.2 Bidirectional RNNs

Considered earlier RNNs have the ability to capture information only from the past $x^{(1)}, .., x^{(t-1)}$ and present $x^{(t)}$ input at time t. However, for certain types of tasks it could be useful to have the output prediction dependent on the whole input sequence. Those tasks include speech recognition and handwriting recognition. The bidirectional RNNs (Schuster and Paliwal, 1997) were invented in that idea in mind. They 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 (Goodfellow, Bengio, and Courville, 2016). Fig.3.3 demonstrates how input sequence is processed by bidirectional RNN.

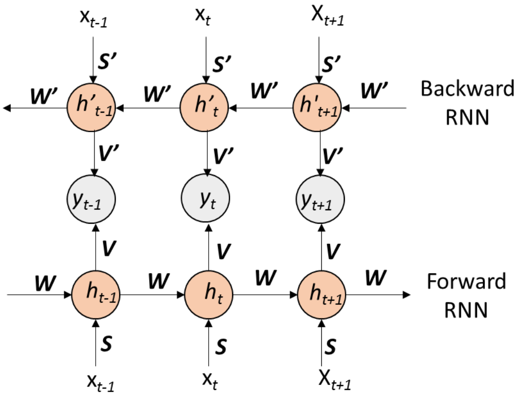
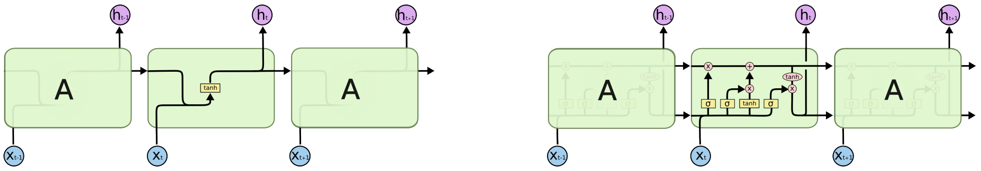


FIGURE 3.3: Bidirectional RNN architecture. [Image source]

3.1.3 Gated RNNs

One ubiquitous problem arises during the training of RNNs - vanishing(or firing) gradients, caused by the need to propagate gradient over many stages - since when unrolled, RNN is as deep as the longest input sequence it processed - to capture long-term dependencies. As of today, the most popular solution to that problem is gated RNN, such as LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Chung



(A) RNN cell [Image source]

(B) LSTM cell [Image source]

FIGURE 3.4: The differences between vanilla RNN cell and LSTM cell

et al., 2014). They are based on the idea of creating paths through time that have derivatives that neither vanish nor explode (Goodfellow, Bengio, and Courville, 2016). Apart from standard hidden state $h^{(t)}$ LSTM cell maintains also cell state $C^{(t)}$, with the possibility to add or remove information from it. And the model basically learns which information from $h^{(t-1)}$ and $C^{(t)}$ store, forget and forward. Fig.3.4 clearly demonstrates the differences between the flow of information in vanilla RNN cell and LSTM cell.

3.2 Seq2Seq architecture

Based on the ability of RNNs to encode an important information from the arbitrary length input sequence in fixed size vector and map that vector to the arbitrary length output sequence seq2seq architecture was created (Sutskever, Vinyals, and Le, 2014; Cho et al., 2014a). As is shown in Fig.3.5 it is composed of two networks: encoder and decoder, both of them are RNNs with the particular architecture, which was discussed in 3.1.1. Encoder encodes the input sequence X into the fixed size context vector $C = h_{enc}^{(T)}$. Then the hidden state of decoder is initialized $h_{dec}^{(0)} = C$. The choice of the output $\hat{y}_t$ is conditioned on the previous output $\hat{y}_{t-1}$ and on the compressed X in the form of dynamically created context vector C. The model is trained by minimizing the log-likelihood between the predicted output $\hat{Y}$ and the target output Y. The teacher forcing method (Lamb et al., 2016) is used during the training process, which suggests using the real target outputs as each next input during training. The alternative is to use the decoder's own guess as the next input. Those two approaches are alternated during the training. During the decoding stage choosing the value with the highest probability doesn't always result in maximum likelihood of the whole output sequence. In order to explore several versions - hypotheses - of the output sequence beam search algorithm (Wiseman and Rush, 2016) is used. It uses

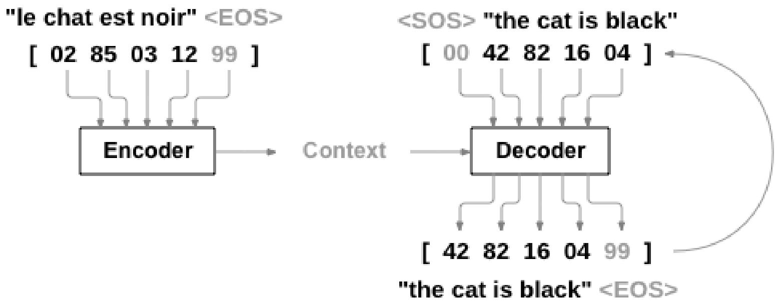


FIGURE 3.5: seq2seq architecture. [Image source]

breadth-first search to build its search tree, but only keeps top k (beam size) nodes (hypotheses with the highest likelihood) at each level(decoding timestep) in memory. The next level will then be expanded from these N nodes. It is still a greedy algorithm, but a lot less greedy than the previous one as its search space is larger.

3.3 Attention mechanism

One clear limitation of the above architecture is when the context vector C has too few dimensions to include all necessary information from the input sequence. (Bahdanau, Cho, and Bengio, 2014) addressed that issue and developed an attention mechanism, which essentially allows decoder to focus on parts of input sequence relevant to output at current timestep .

3.3.1 Global attention

As attention mechanism suggests, context vector c_t is now dynamically created at each decoding timestep (see Fig.3.6). It is computed as a weighted sum of the encoder's hidden states $h_{1:|X|}$:

$$c_t = \sum_{i=1}^{|X|} a_{ti} h_i,$$

whose weights a_t are found by an attention mechanism:

$$a_{ti} = softmax(e_{ti}), e_t = A(\hat{y}_{t-1}, s_{t-1}),$$

where A is an alignment model, implemented as feed-forward neural network of one fully-connected layer, and s_{t-1} is the previous hidden state of the decoder. A is jointly trained with the rest of the model.

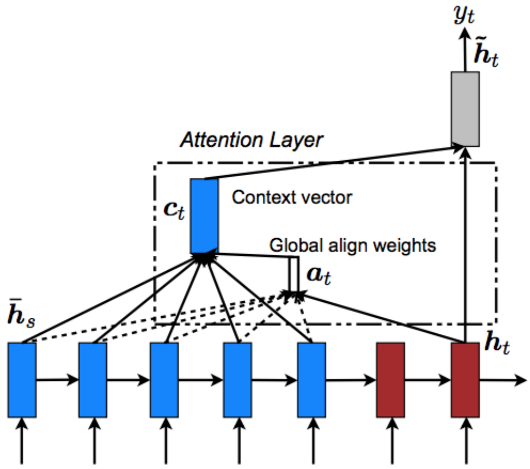


FIGURE 3.6: Global attention in Seq2Seq architecture. [(Luong, Pham, and Manning, 2015)]

3.3.2 Copying mechanism

Yet another challenge in sequence generation is thhandling of the unknown words - words, which are not present in model vocabulary. Input sequences often contain names, places and other specific words, which have to be included in the output sequence, but there is no way for the model to predict them. To tackle this issue (Nisioi et al., 2017b) use the following workaround: when coming across the unknown token in the output sequence, they look at which position in the input sequence has the biggest attention weight when unk was generated and replace it with the word at that position from the input sentence.

More advanced solution was proposed by (Gu et al., 2016). Their model learns to blindly copy some entries from the input sequence to the output at the decoding stage.

3.4 Transfer learning

Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned(Torrey and Shavlik, 2010; Raina et al., 2007). Today it is extremely popular in the field of computer vision, where the weights of the model trained on a gigantic image dataset, for instance ImageNet (Deng et al., 2009), are used to initialize another model with. Then the model is slightly tuned or completely trained for a new task. NLP field could also benefit greatly from such practice, as the models for natural language processing start to require more and more data. For now, one commonly used transfer learning tool in NLP is word embeddings.

3.4.1 Word embeddings

Word embeddings are vectorized representations of words, constructed in such way that the vectors for words which are close in meaning will lay close to one another in the vector space, where the notion of closeness is defined by cosine similarity (Mikolov et al., 2013). Such embeddings are usually learned by a neural network in one of the following ways: CBOW or skip-gram. The CBOW model will learn by trying to predict a word from its context (which could be problematic in case of rare words), while the skip-gram model will learn by predicting a context from a word. Word embeddings are essentially the first attempt to apply transfer learning parading in NLP domain. Pretrained on huge corpuses of data word vectors for large number of languages are available online and incorporating them into one's model substantially improves its performance.

Chapter 4

Proposed solution

4.1 Dataset

4.1.1 Collecting the dataset

After the examination of available sentence simplification datasets, we decided to collect our own dataset, which would contain normal - simple sentence pairs with sufficient degree of simplification, thus helping us to achieve our objectives in sentence simplification. We came up with two ideas of how to collect such a dataset:

- Scrape sites for English learners where the same texts are re-written in different levels of English.
- Exploit the feature of the most modern online news articles: usually the first sentence of the article is a paraphrased and more complex version of the headline.

4.1.2 Dataset characteristics

We scraped 25000 sentence pairs from site and site. They present texts in 3 levels, so we constructed pairs 1-2, 1-3, 2-3 to increase the number of samples. Since from the source we obtain short texts, sentences in them need alignment. We achieved this using dynamic programming and BLEU score - sentence similarity score based on n-grams count (Papineni et al., 2002). However, the matching of sentences is far from perfect, as a result, there are around 15% of completely mismatched sentence pairs (for which chrf score (Popović, 2015) is very small) and one must remove such pairs from the dataset, shrinking its size even more. Table 4.2, Table 4.3 and Table 4.4 show examples of sentence pairs with different English levels. We obtained 30000 sentence pairs from news headlines in articles from news sites such The Washington Post, The Sun, The Guardian and others. We applied basic filtering with chrF score to collected pairs, so not to include sentences which are 'far' from each other in meaning. Table 4.5 shows examples of sentence pairs from news headlines. It is worthy noticing, that the simplified sentences in our dataset were made by exploiting all three necessary sentence simplification components: splitting, deletion and paraphrase, therefore the model has a chance to learn correct sentence simplification in all its fullness.

4.2 Model

Sentence simplification heavily resembles machine translation task, therefore we chosen to experiment with seq2seq encoder-decoder model with global attention mechanism. We also use pretrained Glove word embeddings (Pennington, Socher, and Manning, 2014). We prepared two models: one which architecturally is the same as in (Nisioi et al., 2017b), but trained on Newsela dataset + our dataset(160K sentence pairs); another which architecturally is the same as in (Klein et al., 2017) for text summarisation, which was trained on CNN/Daily Mail dataset (300K articles), as used in (See, Liu, and Manning, 2017). We loaded weights from their model and fine-tuned on our task and our data. Since the size of our sentence simplification dataset is very modest, we speculated that the model could benefit from already learned from text summarization data language model. Further on in text we will refer to the first model as ’from scratch’ and to the second as ’fine-tuned’. Table 4.1 summarises hyperparameters used for training of our models.

4.2.1 Training details

Fine-tuned model was trained in several different settings: froze the encoder and fine-tune the decoder; froze the encoder and train the decoder from scratch; fine-tune both encoder and decoder. For its training we used OpenNMT library (Klein et al., 2017), code for from scratch model[1] was written using PyTorch framework (*PyTorch*), for data processing we used torchtext (*torchtext*).

[1] code is available on github

TABLE 4.1: Models' hyperparameters.

	From scratch model	Fine-tuned model
cell type	LSTM	LSTM
# of layers	2	1
rnn hidden size	256	512
word embeddings size	300	100
bidirectional encoder	-	+
attention type	MLP	MLP
copy attention	-	+
dropout	0.3	0.4
max gradient norm	5	2
learning rate	0.01	0.4
learning rate decay	-	0.3
# of training epochs	14	10
batch size	128	128

TABLE 4.2: Sentence pairs: level 1 and level 2

Normal sentence	Simplified sentence
The zoo is home to around 2,000 animals of around 400 species, and it receives around two million visitors each year.	The zoo is home to 2,000 animals. There are 400 different kinds of animals at the zoo.
The situation is so bad that surfers sometimes have to go around the plastics to surf.	It is so bad that surfers sometimes cannot surf. Two surfers want to make things better.
Three individuals aged 16, 17, and 20 entered a watch shop in Santiago, Chile.	Three men go into a watch shop. They are 16, 17, and 20.
Two million people had orders to evacuate.	Two million people must evacuate.
Recently, there was heavy rainfall in Lviv, a city in the west of Ukraine.	Lviv is a city in the west of Ukraine. It rains heavily there.

TABLE 4.3: Sentence pairs: level 1 and level 3

Normal sentence	Simplified sentence
The world's throwaway culture is slowly turning the Galapagos Islands into a plastic wasteland, as plastics from South America and the Pacific are littering the beaches of the archipelago with plastic bottles being the most common item.	The Galapagos Islands are in the Pacific Ocean. Plastics from South America and the Pacific are littering the islands. The Galapagos Islands are home to hundreds of endemic species.
Ten-month-old miniature horse Honor visited young patients well enough to meet him in person at a hospital in New York.	A miniature horse visits a hospital in New York.
In Britain, Alf Smith turns 110 years old and calls another man named Bob Weighton, who has also just turned 110 years old.	In Britain, two men turn 110 years old.
The Attorney General for New South Wales, Australia, believes that less anxiety means more reliable evidence, so dogs are coming to law courts in Sydney to reduce stress.	People bring dogs to law courts in Sydney, Australia.
This Chinese New Year celebrates the Year of the Dog as based on the traditional Chinese zodiac.	This Chinese New Year celebrates the Year of the Dog. The animal comes from the Chinese zodiac.

TABLE 4.4: Sentence pairs: level 2 and level 3

Normal sentence	Simplified sentence
The popular snack, rich in protein and vitamins, is becoming harder to find in Asia and the price has gone up.	Since they are so popular, their price is also up. The spiders are rich in protein and vitamins.
Jodie Bradshaw, who runs FreeHearts Animal Sanctuary in Tasmania, spotted a dead wombat by the roadside.	A woman saw a dead wombat by the road in Tasmania.
The four-year-old child was dangling on a balcony on a fourth floor and would have certainly fallen down and died if he had not been saved.	A four-year-old boy dangled from a balcony on a fourth floor in Paris. Luckily, Mamoudou Gassama saved him.
Heavy rainstorms have hit parts of eastern China's Anhui and Jiangxi provinces, leaving residents of several regions trapped and crops destroyed.	Heavy rainstorms hit parts of eastern China. Water trapped people in several regions and destroyed crops.
Anwar Ibrahim is Malaysia's reformist icon, and police released him from jail after the prime minister asked for a royal pardon.	Anwar Ibrahim left jail after the prime minister asked for a royal pardon.

TABLE 4.5: Sentence pairs: news headline and first sentence of the article

Normal sentence	Simplified sentence
KABUL - An investigation into a November firefight between Taliban insurgents and joint U.S. and Afghan forces- has concluded that 33 civilians were killed in the operation, the U.S. military said Thursday.	U.S. military says battle with Taliban killed 33 civilians in Afghanistan.
A Connecticut man who jumped the White House fence draped in an American flag on Thanksgiving Day in 2015 was sentenced Thursday to three years in probation and ordered to stay away during that time from the District and people and places under U.S. Secret Service protection.	Conn. man who jumped White House fence draped in U.S. flag sentenced to 3 yrs. probation.
A skinny sea lion pup with a taste for the finer things in life turned up on Thursday inside a fancy San Diego, California, restaurant, where it settled down into a booth.	Hungry sea lion looks for meal at San Diego restaurant.
A former Baltimore police officer acquitted of animal cruelty charges after he slit a dog's throat will receive $45,000 in back pay from city government.	Former officer who cut dog's throat to get $45,000 in back pay.
Elizabeth Strohfus, who piloted military planes across the country during World War II and received two Congressional Gold Medals, died March 6 at an assisted-living center in Faribault, Minn. She was 96.	Elizabeth Strohfus, World War II-era pilot, dies at 96.

Chapter 5

Results

5.1 Experiments results

Firstly, it has to be noted that the existing sentence simplification metrics (i.e SARI (Xu et al., 2016)) or widely used BLEU score (Papineni et al., 2002), which in fact was developed for machine translation evaluation, don't exactly measure the quality of sentence simplification, they only evaluate the similarity of sentences. In our case this is a little useless because:

- For almost every complex sentence there exists several ways to simplify it. So if the predicted simplified sentence is has low BLEU score with the target simple sentence, it doesn't necessarily mean the simplification was bad.
- The predicted simple sentence can be close in meaning to the source complex sentence, but still missing an important piece of information, which renders the simplification useless.

Therefore, we present our results in form of a table with hand-picked from test set target simple sentences and predicted simplifications for human judgement. Rule-based simplifications are generated by (Cetto et al., 2018). For baseline models we use (Nisioi et al., 2017b) model, trained on 280K sentence pairs(falling under *good matches* and *partial matches* categories) from Parallel Wikipedia Corpus (Coster and Kauchak, 2011). Analysing the results in Table.5.1, one can eagerly observe:

- Models trained on sentence simplification dataset act more like autoencoders, making almost no changes in the sentence. Apparently the dataset is too small and diverse to learn some solid simplification patterns from scratch.
- Since the beam search is used in baseline and from scratch models to choose the output sequence with the highest probability, sometimes the word at the end is repeated.
- Fine-tuned model produces very short sentences, because it was initially trained for a few words text summarization.
- Sentences produced by fine-tuned models often does not make sense, as if the model hasn't been trained enough. However, at 12th epoch, the validation metrics stop improving, while metrics on train are still changing - model is overfitting. Apparently, it isn't that easy to transfer knowledge from one task to another.
- Unfortunately, this results cannot yet beat the rule-based sentence simplification for question generation. Table continues in A.

TABLE 5.1: Comparison of sentence simplifications performed by different models. Continues in A

Original sentence	In celebration of World Puppy Day, this super talented pooch, Purin, has set a new Guinness World Record.
Target simplified sentence	In 2015, a dog sets a world record.
Rule-based simplification	Purin is this super talented pooch. This super talented pooch has set a new Guinness World Record. This is in celebration of World Puppy Day.
Baseline simplification	In celebration of World Puppy Day, this super talented pooch, has set a new Guinness World Record. this .
Model from scratch simplification	This super talented pooch, Purin, has set a new Guinness World Record.
Fine-tuned model simplification	In talented pooch, Purin, sets new Guinness World Record.
Original sentence	In 2014, a park employee was also mauled to death by a tiger
Target simplified sentence	A tiger kills a person who works at the park
Rule-based simplification	A park employee was also mauled to death by a tiger. This was in 2014.
Baseline simplification	In 2014, a park employee was also mauled to death by a tiger.
Model from scratch simplification	In 2014, a park employee was also killed to death by a tiger
Fine-tuned model simplification	In park worker dies at a park tiger is fine
Original sentence	The publishing company which owns The New Yorker, Vanity Fair and other magazines, is the first renter to set up offices in the tower where it will occupy floors 20 to 44.
Target simplified sentence	One company moves into the building. It has floors 20 to 44. More people will move in the building in 2015
Rule-based simplification	Is the first renter to set up offices in the tower where it will occupy floors 20. This is the publishing company which owns The New Yorker, Vanity Fair and other magazines. This is to 44.
Baseline simplification	The publishing company which owns The New Yorker, Vanity Fair and other magazines, is the first renter to set up offices in the tower where it will occupy floors 20 to 44. 44. .
Model from scratch simplification	The company which owns the New Yorker, Vanity Fair and other magazines. is the first renter to set up offices in the tower where it will occupy floors 20 to 44.
Fine-tuned model simplification	The is the first renter to set up offices in the tower
Original sentence	A brother and a sister burned to death as they watched the eruption from a bridge
Target simplified sentence	They are a brother and a sister. This eruption is deadly.
Rule-based simplification	A brother and a sister burned to death. This was as they watched the eruption from a bridge.
Baseline simplification	A brother and a sister burned to death .
Model from scratch simplification	A brother and a sister burns to death as they watched the eruption from a bridge.
Fine-tuned model simplification	A is from a bridge.

5.2 Conclusion

5.2.1 Contribution

We presented an extension to the existing sentence simplification dataset and experimented with transfer of knowledge from text summarization to sentence simplification.

5.2.2 Possible improvements and future work

We see possible room for improvements in:

1. Dataset: design a more effective sentences alignment method; develop a complex metric to evaluate the quality of sentence simplification.

2. Model: look into bigger pretrained models for NLP, for example, here; explore the nuances of transfer learning in case of text data; combine the paraphrasing potential of the deep learning models and the syntactic rules for simplification.

Appendix A

Examples of simplified sentences

TABLE A.1: Comparison of sentence simplifications performed by different models

Original sentence	People must turn to the black market where products, like milk or sugar, are 10 times more expensive.
Target simplified sentence	People must turn to the black market.Some things are 10 times more expensive there.
Rule-based simplification	People must turn to the black market where are 10 times more expensive. This is like milk or sugar. This is products.
Baseline simplification	People must go to the black market where products, like milk or sugar, are 10 times more expensive.
Model from scratch simplification	People must turn to the black market where products, like milk or Trucks are 10 times more expensive.
Fine-tuned model simplification	People must turn to the black market where products, like milk or sugar,
Original sentence	Priscilla said that her husband was a giving man and he would have liked sharing the things that he accomplished with his fans.
Target simplified sentence	Priscilla thinks that Elvis would like this. He liked sharing things with his fans.
Rule-based simplification	Her husband was a giving man. This was what Priscilla said. And he would have liked sharing the things that he accomplished with his fans.
Baseline simplification	Priscilla said that her husband was a giving men and he would have liked sharing the things that he accomplished with his fans.
Model from scratch simplification	Priscilla said that her husband was a giving man. He would have liked sharing the things that he accomplished with his fans.
Fine-tuned model simplification	Priscilla is a giving man
Original sentence	This fish blows itself up like a balloon when it is in danger.
Target simplified sentence	The fish is very special. It gets big when it is in danger.
Rule-based simplification	This fish blows itself up like a balloon when it is in danger.
Baseline simplification	This fish blows itself up .
Model from scratch simplification	This fish blows up like a balloon. It is in danger.
Fine-tuned model simplification	This fish blows itself when it is in danger.
Original sentence	Ten per cent of people in the UK no longer have cash in their pockets, and one third of people have less than five pounds.
Target simplified sentence	Britain is one of them. Ten per cent of people in Britain have no cash in their pockets.
Rule-based simplification	Ten per cent of people in the UK no longer have cash in their pockets. And one third of people have less than five pounds.
Baseline simplification	Ten per cent of people in the UK no longer have cash in their pockets, .
Model from scratch simplification	en per cent of people in the UK no longer have cash in their pockets. One third of people have less than five pounds.
Fine-tuned model simplification	Ten cent of people have less than five pounds.
Original sentence	Renault is going to unveil an electric car that is actually also a sports car.
Target simplified sentence	Renault shows an electric sports car.
Rule-based simplification	Renault is going to unveil an electric car that is actually also a sports car .
Baseline simplification	Renault is going to unveil an electric car.
Model from scratch simplification	Renault is going to unveil an electric car. It is actually a sports car.
Fine-tuned model simplification	Renault is also a sports car.

Bibliography

Adams, Douglas (1982). *Life, the Universe and Everything*. 1st ed. Vol. 2. 6. Pan Books (UK). ISBN: 9788804391302.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473*.

Carroll, John A. et al. (1999). “Simplifying Text for Language-Impaired Readers”. In: *EACL*.

Cetto, Matthias et al. (2018). “Graphene: Semantically-Linked Propositions in Open Information Extraction”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2300–2311. URL: http://aclweb.org/anthology/C18-1195.

Chandrasekar, Raman, Christine Doran, and Srinivas Bangalore (1996). “Motivations and Methods for Text Simplification”. In: *COLING*.

Chen, Danqi and Christopher D. Manning (2014). “A Fast and Accurate Dependency Parser using Neural Networks”. In: *EMNLP*.

Cho, Kyunghyun et al. (2014a). “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078*.

Cho, Kyunghyun et al. (2014b). “On the properties of neural machine translation: Encoder-decoder approaches”. In: *arXiv preprint arXiv:1409.1259*.

Chung, Junyoung et al. (2014). “Empirical evaluation of gated recurrent neural networks on sequence modeling”. In: *arXiv preprint arXiv:1412.3555*.

Coster, William and David Kauchak (2011). “Simple English Wikipedia: A New Text Simplification Task”. In: *ACL*.

Csáji, Balázs Csanád. “Approximation with artificial neural networks”. In:

Deng, J. et al. (2009). “ImageNet: A Large-Scale Hierarchical Image Database”. In: *CVPR09*.

Devlin, Siobhan (1999). “Simplifying natural lanauge text for aphasic readers”. In:

Evans, Richard, Constantin Orasan, and Iustin Dornescu (2014). “An evaluation of syntactic simplification rules for people with autism”. In: *PITR@EACL*.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch (2013). “PPDB: The Paraphrase Database”. In: *HLT-NAACL*.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Gu, Jiatao et al. (2016). “Incorporating copying mechanism in sequence-to-sequence learning”. In: *arXiv preprint arXiv:1603.06393*.

Heilman, Michael (2009). “Question Generation via Overgenerating Transformations and Ranking”. In:

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.

Inui, Kentaro et al. (2003). “Text Simplification for Reading Assistance: A Project Note”. In: *IWP@ACL*.

Kaji, Nobuhiro et al. (2002). “Verb Paraphrase based on Case Frame Alignment”. In: *ACL*.

Klebanov, Beata Beigman, Kevin Knight, and Daniel Marcu (2004). "Text Simplification for Information-Seeking Applications". In: *CoopIS/DOA/ODBASE*.

Klein, Guillaume et al. (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proc. ACL*. DOI: 10.18653/v1/P17-4012. URL: https://doi.org/10.18653/v1/P17-4012.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu (2003). "Statistical Phrase-Based Translation". In: *HLT-NAACL*.

Lamb, Alex M et al. (2016). "Professor forcing: A new algorithm for training recurrent networks". In: *Advances In Neural Information Processing Systems*, pp. 4601–4609.

Levy, Roger and Galen Andrew (2006). "Tregex and Tsurgeon: tools for querying and manipulating tree data structures". In: *LREC*.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning (2015). "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025*.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *NIPS*.

Nisioi, Sergiu et al. (2017a). "Exploring Neural Text Simplification Models". In: *ACL*.

Nisioi, Sergiu et al. (2017b). "Exploring Neural Text Simplification Models". In: *ACL (2)*. The Association for Computational Linguistics.

open-source. *PyTorch*. Version 1.1.0. URL: https://pytorch.org.

— *torchtext*. Version 0.4.0. URL: https://github.com/pytorch/text.

Papineni, Kishore et al. (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 311–318.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Popović, Maja (2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395.

Raina, Rajat et al. (2007). "Self-taught Learning: Transfer Learning from Unlabeled Data". In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvalis, Oregon, USA: ACM, pp. 759–766. ISBN: 978-1-59593-793-3. DOI: 10.1145/1273496.1273592. URL: http://doi.acm.org/10.1145/1273496.1273592.

Rello, Luz et al. (2013). "DysWebxia 2.0!: more accessible text for people with dyslexia". In: *W4A*.

Rumelhart, David E, Geoffrey E Hinton, Ronald J Williams, et al. "Learning representations by back-propagating errors". In: *Cognitive modeling* 5.3, p. 1.

Schuster, Mike and Kuldip K Paliwal (1997). "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11, pp. 2673–2681.

See, Abigail, Peter J Liu, and Christopher D Manning (2017). "Get to the point: Summarization with pointer-generator networks". In: *arXiv preprint arXiv:1704.04368*.

Smith, David A. and Jason Eisner (2006). "Quasi-Synchronous Grammars: Alignment by Soft Projection of Syntactic Dependencies". In: *WMT@HLT-NAACL*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*, pp. 3104–3112.

Torrey, Lisa and Jude Shavlik (2010). "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global, pp. 242–264.

Vickrey, David and Daphne Koller (2008). "Sentence Simplification for Semantic Role Labeling". In: *ACL*.

Vu, Tu et al. (2018). “Sentence Simplification with Memory-Augmented Neural Networks”. In: *NAACL-HLT*.

Watanabe, Willian Massami et al. (2009). “Facilita: reading assistance for low-literacy readers”. In: *SIGDOC*.

Wiseman, Sam and Alexander M Rush (2016). “Sequence-to-sequence learning as beam-search optimization”. In: *arXiv preprint arXiv:1606.02960*.

Woodsend, Kristian and Mirella Lapata (2014). “Text Rewriting Improves Semantic Role Labeling”. In: *J. Artif. Intell. Res.* 51, pp. 133–164.

Wubben, Sander, Antal van den Bosch, and Emiel Krahmer (2012). “Sentence Simplification by Monolingual Machine Translation”. In: *ACL*.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles (2015). “Problems in current text simplification research: New data can help”. In: *Transactions of the Association for Computational Linguistics* 3, pp. 283–297.

Xu, Wei et al. (2016). “Optimizing statistical machine translation for text simplification”. In: *Transactions of the Association for Computational Linguistics* 4, pp. 401–415.

Yamada, Kenji and Kevin Knight (2001). “A Syntax-based Statistical Translation Model”. In: *ACL*.

Zhang, Xingxing and Mirella Lapata (2017). “Sentence Simplification with Deep Reinforcement Learning”. In: *EMNLP*.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych (2010). “A Monolingual Tree-based Translation Model for Sentence Simplification”. In: *COLING*.