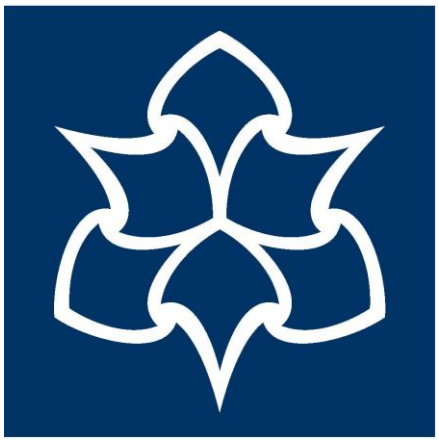
COMBINMT: AN EXPLORATION INTO NEURAL TEXT SIMPLIFCATION METHODS

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY FOR THE DEGREE OF MASTER OF SCIENCE

IN THE FACULTY OF SCIENCE AND ENGINEERING



2019

By

Michael James Cooper
Department of Computing and Mathematics

Contents

Abstract

We present a replication study of *Exploring Neural Text Simplification Models* (Nisioi et al., 2017), our first look into machine learning. By replicating the methods presented within this paper, we found that we gained similar results.

We used a different implementation of OpenNMT, and incorporated the Newsela corpus alongside the Hwang et al., (2016) Wikipedia dataset. After running evaluations on the different datasets to ensure they were of a high quality, we reduced them further by means of cosine similarity, and trained a model on this new, combined dataset, and used locally trained embeddings and pretrained Google News vectors. The resulting systems were dubbed combiNMT, one being combiNMT995, had a cosine similarity cut off at anything over 0.995, and combiNMT98, which had a cosine similarity cut off at anything over 0.98.We then performed an extended version of the human evaluation used by the original research.

With the extended human evaluation showing our system performs better than previous models in terms of correct changes to text, it also performs well in terms of grammaticality and meaning preservation. The system performs comparatively with the NTS systems by way of SARI scores.

As far as the NTS systems performance, this study finds that they do indeed perform well when ranked on the right metrics.

A complete copy of the study can be found at : https://github.com/mgnc2867/combiNMT

Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work.

Signed: 

(Michael James Cooper)

Date: 27/09/2019

Acknowledgements

There are a few people who have helped me, either knowingly or unknowingly in the process of this research.

Firstly, this final paper would have been unimaginable without the help provided by Matthew Shardlow. His guiding hand has been much appreciated, as has his critical eye.

Secondly, my parents, Dave and Jeannette Cooper who've assisted in many ways, mostly by listening to me airing my thoughts, even though they had no idea what I was talking about.

Thirdly, Ashlee Cox – for much needed head space.

Fourthly, Jake Roberts, for the unending stream of encouragement, and for providing me with quiet working space.

Fifthly, Gus and Chuck – for the constant affection.

Abbreviations

NMT – Neural Machine Translation
NLP – Natural Language Processing
RNN – Recurrent Neural Networks
BLEU – Bi-Lingual Evaluation Understudy
GLUE – General Language Understanding Evaluation
RBMT – Rule Based Statistic Machine Translation
NPMT – Neural Phrase Based Machine Translation
SMT – Statistical Machine Translation
S2S – Sequence to Sequence model
TS – Text Simplification
NTS – Neural Text Simplification model
LSTM – Long Short Term Memory
DNN – Deep Neural Networks
ATS – Automated Text Simplification
SGD – Stochastic Gradient Descent
MT – Machine Translation
MLM – Masked Language Model

Chapter 1

Introduction

Neural Machine Translation (NMT) is a recent development which has brought a lot of attention to the field of Natural Language Processing (NLP.) NMT uses Recurrent Neural Networks (RNNs) to imitate the neurons in the human brain making connections, learning new information, and presents the capability of accessing the input information in its entirety rather than just part by part. NMT scores more highly on standard baseline metrics, such as Bi-Lingual Evaluation Understudy (BLEU) score (Papenini et al., 2002) and the General Language Understanding Evaluation (GLUE) benchmark leaderboard (Wang et al., 2018).

NMT has brought around significant improvements in the field, especially when compared to Rule Based Statistic Machine Translation (RBMT) (Weiss et al,1995,) Neural Phrase Based Machine Translation (NPMT) (Huang et al., 2017) and Statistical Machine Translation (SMT) systems (Wu et al., 2016.) Since its conception, using purely sequence-to-sequence (S2S) models (Sutskever et al., 2014, Cho et al., 2014,) NMT has become a widely used technique in machine translation, as well as a well-regarded approach for other tasks including dialogue generation, parsing and summarization.

This paper is concerned primarily with a specific subsection of translation tasks, namely text simplification. Text simplification (TS) effectively works like any other Machine Translation Task, except it uses a simplified version of the original language as its output language, rather than a different language. TS is the act of reducing the complexity of text for people with lower comprehensive level, due to low literacy, aphasia or learning difficulties. There are three main measures to be mindful of when working with Text Simplification. These are grammaticality, meaning preservation, and simplicity. The output from the system should be grammatically correct, retain its original meaning and be simpler to understand than the input.

Along with the attention NMT has brought to NLP, there have also been advances in the technology surrounding the tasks. The research in this paper has used several of the most highly regarded, as well as some of the most recent developments, OpenNMT (Klein et al., 2017) has

been used as the foundation of the system. We have used RoBERTa (Liu et al., 2019) to measure the whether the simplified sentences are a simpler version of the original sentence, whether the simplified version makes sense when placed immediately after the original (i.e. entailment) or whether the two sentences are contradictions of each other. We used the Python library NumPy (Oliphant., 2006) to compute the cosine similarity between the original and simple sentences which RoBERTa suggested as simplified versions. From there, we deleted the most similar sentences to ensure that the system learnt something from the difference in the sentences.

A lot of the advances in the field have been made by big companies such as Google (who Ilya Sutskever was working for when he presented the S2S model (Sutskever et al., 2014)) and Facebook (who Yinhan Liu is working for whilst working on the RoBERTa model. (Liu et al., 2019)) Although not necessarily true of all big companies, Google and Facebook seem to release the research they've conducted quite freely, though not with unrestricted licences. All publicly presented models of RoBERTa are available on GitHub quite openly to the public. This ensures smaller research groups are not excluded from the possibilities of progressing the field. With the multitude of translation tasks within the field, open source software ensures growth for all, not only for the task for which it was originally designed. For example, RoBERTa was designed to ensure the quality of Facebook AI's chatbots, but also achieved an impressive score on the GLUE leaderboard, and state-of-the-art results on 4 out of 9 GLUE tasks, including the Semantic Textual Similarity Benchmark (STS-B) which we have used it for.

The aims of this research has been to work towards replicating the Neural Text Simplification (NTS) system presented in *Exploring Neural Text simplification Models* (Nisioi et al., 2017), as well as their results, based on the same metrics, as part of the shared task for International Conference on Language Resources and Evaluation (LREC) 2020.

The objective of this research, other than this paper, and the literary review contained within, is to submit a paper to the LREC 2020 conference in Marseille describing the use of Nisioi et al.'s system, the ease and difficulties of replicating the results, and laying out any improvements made in comparison to the original results.

There will be an extended literature review which covers the technologies discussed above, as well as some other important developments in the field over recent years. Proceeding from this, the paper will lay out how these technologies were used in the system presented and the

configurations used. The results from a human evaluation process, as well as against the standard translation metrics, BLEU and measuring System output Against References and against Input sentence (SARI), will then be laid out, with an evaluation of the results following. Then, finally, a conclusion, which will also discuss potential further research. Extended examples of the results will be included in the appendices.

Chapter 2

Related Works

This section will focus on technologies either used in this research or with a very close relation to the subject matter. This ranges from well-established technologies such as RNNs to more recent developments, such as RoBERTa, released in July 2019. It hopes to provide a more in depth overview of the required technologies.

Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) aimed to solve the problem arising in backpropagation of 'error signals "flowing backwards in time."' By using a Recurrent Neural Network (RNN) in 'in conjuncture with an appropriated gradient-based learning algorithm', LSTM was designed to remember inputs for much longer than other RNN systems without losing the capabilities it possesses over a short time lag.

The advantages of LSTM include its ability to bridge long time longer times, whilst being able to handle unstructured text data within structured or semi-structured text data. It doesn't require a pre-defined choice of a finite number of states, in theory, it can deal with any number of states. States, in this case, refers to the activation of the neuron within any given memory unit. The activation within the neuron a combination of previous inputs. At each training step, the memory unit is fed input, it then decides how much of the input to fed to the neuron itself, how much of its previous activation to keep and how much of the activation to output.

Generating Sequences with Recurrent Neural Networks (Graves, 2014), states that 'by making the network treat its inventions as if they were real, much like a person dreaming,' novel sequences can be generated by sampling from the output, and feeding that sample as input at the next step. Due to their use of locally trained embeddings, to better learn the differences between training examples, RNNS are distinguished from other models. This is down to their ability to interpret and represent the training data in a complex way, and rarely present identical outputs. When used as LSTM, Graves states that RNNs can use the memory capabilities to 'generate complex, realistic sequences containing long-range structure.'

Sequence to Sequence Learning with Neural Networks, (Sutskever et al., 2014) presents a new end-to-end approach which makes 'minimal assumptions' on the sequence structure. This is an improvement on the Deep Neural Networks (DNNs) used extensively until this point. DNNs

are ‘extremely powerful’ models that excel on problems such as speech recognition and object recognition. However, despite their flexibility they can only be used when the input and output sequences of the problems can be mapped to vectors with specific length, reducing the amount of ‘important problems’ DNNs can tackle due to the ‘long-term dependencies’ on the data structure.

By ‘extracting translations from an ensemble of 5 deep LSTMs, the researchers improved, over the entire Workshop on Statistical Machine Translation 2014 (WMT’14) dataset (Association of Computational Linguistics, 2014,) on the state-of-the-art Bi-Lingual Evaluation Understudy (BLEU) (Papineni et al., 2002) score on the English to French translation tasks. They also discovered that reversing the source sentences causes the LSTM to ‘learn much better.’ The system perplexity dropped from 5.8 to 4.7, and the BLEU score improved by almost 5 points. The research could not provide a full explanation of this phenomenon, but ‘believe it is caused by the introduction of many short-term dependencies.’ The improvement could also be due to a reduction in the ‘minimal time lag;’ when source sentences are concatenated with target sentences, each word from the source sentence is far from its corresponding word in the target sentence, but by reversing the order of the source sentences, although the average distance is unchanged, the first few words in the source language are now very close to the first few words in the target language. They argue that backpropagation has an easier time ‘establishing communication between the source sentence and the target sentence, […] substantially improv[ing] overall performance.’ There is, however, also the possibility that this phenomenon could be down to an unstable system. Different permutations during training could lead an unstable system to produce vastly different results when compared on automated metrics.

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (Cho et al., 2014) focused on an RNN Encoder-Decoder pair, that is two Recurrent Neural Networks working alongside one another, one acting as an encoder, mapping a sentence of variable length to a vector of fixed length, the other as a decoder, mapping the representation to a translated sentence of variable length. These two RNNs are ‘trained jointly to maximize the conditional probability of the target sequence given a source sequence.’ They also proposed a sophisticated hidden unit which acted to improve the ease of training, as well as the capacity of the memory. This was trained on translating from English to French, and then used in a phrase based SMT (PBSMT) system which proved an improvement in terms of BLEU score on translation performance.

In the seminal paper *Neural Machine Translation by Jointly Learning to Align and Translate* (Bahdanau et al., 2015), it was also conjectured that fixed-length vectors widely used in Neural Machine Translation (NMT) slow any improvement in performance. Unlike phrase-based translation systems consisting of small separately tuned sub-components, NMT consists of one large neural network, which reads a sentence and then outputs a translated version. Most models proposed before this paper's release used 'an encoder and a decoder for each language or involve[d] a language-specific encoder' the output of which was then compared.

However, they present a solution; an extension to the standard encoder-decoder architecture that learns to both align and translate words within the model before searching for the position in the context vectors where the most relevant information is concentrated. The system then predicts a word based on the context vectors, the source positions and all previously generated words.

This system significantly outperformed the conventional encoder-decoder system, on the same English to French translation task as Sustkever et al, achieving comparable results to most state-of-the-art PBSMT system. 'A striking result, considering the proposed architecture […] has only been proposed as recently as [2014].'

To assist the learning algorithms to achieve better performance, distributed representations of words are commonly used. This is not new technology, it dates to (Rumelhart, Hinton and Williams, 1986). It's been applied to statistical language modelling, and Machine Translation (MT), as well as a large range of Natural Language Processing (NLP) tasks. Introduced in *Efficient Estimation of Word Representations in Vector Space* (Mikolov et al., 2013) Word2Vec embeddings aimed to improve on the quality and training speed of the standard Skip-gram model presented by Mikolov.

By 'sub-sampling of frequent words' in the training phase, the research shows a 'significant' increase in speed, improved accuracy of representation of less frequent words, and better representation of frequently used words. Another interesting result shows that word vectors can be 'meaningfully combined,' this would mean that learnt phrases could then be represented using a single vector token. This gives the advantage of a 'powerful yet simple way to represent longer pieces of text, while having minimal computational complexity.'

Although many models join their Encoder-Decoder pairs through an attention mechanism, the Transformer, a sequence transduction model presented in *Attention Is All You Need* (Vaswani et al, 2017), is an architecture which relies entirely on multi-headed self-attention mechanisms

to draw global dependencies between the input and output sequences, whilst dispensing with the need for recurrence and convolutions entirely. This allows for higher quality translation after being trained 'for as little as twelve hours on eight P100 GPUs.'

Instead of mapping the input to a fixed length vector, the Transformer maps the 'input sequence of symbol representations […] to a sequence of continuous representations.' This is then fed to the decoder which 'generates an output sequence […] of symbols one element at a time.' The transformer is auto-regressive, that is, it feeds the generated symbols as additional input whilst generating the next sequence of symbols. The model was tested on both the WMT 14 English to German (Association for Computational Linguistics, 2014) and WMT 2014 English to French (Association for Computational Linguistics, 2014) translation tasks, achieving new state-of-the-art results. In the English to German task, the best performing Transformer model 'outperforms even all previously reported ensembles.'

There have been multiple advances in the technology used to create vector space representations of words for training within the past 18 months. In *Deep contextualized word representations* (Peters et al., 2018) the researchers assert that 'high quality representations can be challenging,' in that they should be able to represent different syntactic and semantic word use, and yet model how these different uses are used across varied linguistic contexts. This is the challenge which ELMo (Embeddings from Language Models) is based.

To address these issues, they assign each token a 'representation that is a function of the entire input sentence.' They are a representation of all the internal layers of the bidirectional LSTM 'trained with a coupled language model' with which they are created. The use of this 'deep' representation means that they capture both the context-dependent aspects, such as word meaning, which the high-level states of LSTM produce, as well as the syntactic representation created by the low-level states. These representations improved on the state-of-the-art results 'in every considered case across a range of challenging language understanding problems.'

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Devlin et al., 2019) caused a lot of excitement within the NLP community. Bidirection Encoder Representations from Transformers (BERT), the model presented in this paper, is 'conceptually simple and empirically powerful' advancing even further on 11 NLP tasks by significant margins, on both sentence and token level.

BERT uses a Masked Language Model (MLM) which 'randomly masks some of the tokens from the input' with the intention that the model should correctly predict the original

vocabulary based purely on context. It also uses 'next sentence prediction' which 'jointly pretrains text-pair representations.' This research shows how effective bidirectional pretraining can be for language representation models, and how these representation models reduce the need for many heavily-engineered task-specific architectures.

Another development worth noting is *XLNet*: *Generalized Autoregressive Pretraining for Language Understanding* (Yang et al., 2019). This paper claims that by 'relying on corrupting the input with masks, BERT neglects dependency between the masked positions' and therefore suffers from a discrepancy. XLNet 'enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order' whilst, due to its autoregressive nature, the model helps negate the limitations of BERT, building once more on the state-of-the-art results on 18 benchmark tasks '[u]nder a set of fair comparison experiments'.

Research coming out of Facebook AI in July 2019 builds upon the advances presented by BERT. *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Liu et al., 2019) is a replication study of BERT pretraining, which aimed to analyse the effects of the size of training data and many important hyperparameters. It proposes 4 modifications to the way in which BERT is trained. They suggest that training the model with (1) more data, bigger batch sizes and for longer, (2) training on longer sequences, (3) changing the mask pattern, which is applied to the training data, (4) and removing the next sentence prediction task.

By implementing these changes, RoBERTa matches the GLUE score presented by the XLNet team in June on the public leader board. The model also claims state-of-the-art results on 4 of the 9 GLUE tasks, including the Semantic Textual Similarity Benchmark (STS-B).

OpenNMT: Open-Source Toolkit for Neural Machine Translation presented by Klein et al. in 2017 is a greatly important paper. It calls for the NLP community to build on its example by releasing more such open source toolkits, since systems developed by industry are not likely to be released with unrestricted licences, whilst many others only exist as research code. It claims that '[a] toolkit should aim to provide a shared framework for developing and comparing open-source systems, while at the same time being efficient and accurate enough to be used in production context.'

The technology it presents is also extremely impressive. Aiming to build upon the strengths of the *Nematus* system presented by the University of Edinburgh, and striving to provide additional documentation and functionality, Klein et al. set their focus on fast training and

efficiency, modularity and readability and significant research extensibility, additionally, OpenNMT is a complete open source NMT implementation. Initially, it was implemented using the Lua/Torch mathematic framework, due to its ease being 'extended using Torch's internal standard neural network components.' It was extended by Adam Lerer of Facebook Research to support Python/PyTorch framework.

Using an attention-based Encoder-Decoder architecture, as described in (Bahdanau et al., 2014), NMT models the probability of a target sentence, given a source sentence. OpenNMT is a complete library, including vanilla NMT models, along with 'support for attention, gating, stacking, input feeding, regularization, beam search and all other options necessary for state-of-the-art performance.'

Chapter 3

Methodology

This section describes the paper which we attempted to replicate. It goes on to describe the technical aspects of our systems set up. After this are the details of our evaluation process, and finally, our datasets.

3.1 Background

This paper is a replication study of *Exploring Neural Text Simplification Models* (Nisioi et al., 2017), in which the research team presented an Automated Text Simplification (ATS) system which 'addresses the applicability of Neural sequence to sequence models for ATS'. ATS systems are designed to 'transform original texts into different (simpler) variants which would be understood by wider audiences and more successfully processed by various NLP tools.' By making use of advances is NMT, the researchers adapted existing architectures for their task.

The resulting system was named the Neural Text Simplification (NTS) model which used the OpenNMT framework, discussed earlier in this paper, to train and build an architecture with two LSTM layers. They had an RNN Encoder-Decoder pair, connected by an attention mechanism layer, the RNNs had 'hidden states of size 500 and 500 hidden units.' The first of these figures is the number of features the LSTM is able to store at any given time, meaning that by the time a sentence is fed in token vector by token vector, the original token vector should not be forgotten. The second figure is the number of hidden units within each RNN.

In an attempt to reduce the likelihood of the system learning idiosyncratic errors due to the size of the dataset (usually because the dataset is too small), known as Overfitting, NTS has a dropout probability rate of 0.3. Dropout is a regularisation method which, in effect, randomly kills the connections of a node, whilst keeping each connection within the probability rate (0.3 in the case of this study) The probability is that of the token represented by the vector which the node holds. If the node is dropped out, the input and output from that node are ignored as well, resulting in a network of a temporarily reduced size.

The researchers trained the model for 15 epochs over the data, with plain Stochastic Gradient Descent (SGD) optimization and the vocabulary size set to 50,000. An epoch in this case is the number of training steps it takes to make one pass over the data. This is calculated by dividing the number of inputs to be processed by the number of batches. In the NTS system, there were 284,677 sentences run over a batch size of 64, which is the default number on OpenNMT. SGD is a optimization algorithm, one of several which are increasingly popular within the field. These algorithms typically perform unnecessary computations for large datasets, due to the fact it 'recomputes gradients for similar examples before each parameter update' (Ruder, 2017). However, SGD performs one update at a time, doing away with the redundancy.

After the 8th pass over the data, the learning rate of the system is halved. The learning rate is a configurable hyperparameter, which dictates the amount of change to a model during the discovery of the 'weights' of the neural network. The learning rate has a small value, usually between 0.0 and 1.0, however in this study, the parameter is set over 'uniform distribution with support [-0.1, 0.1]' meaning each outcome is initially equally likely At the end of each pass, the state of the model is saved.

On top of this architecture, the researchers employed 'global attention in combination with input feeding' for the decoder. Input feeding in this case is the approach of concatenating the representation of the previous output with the context vector of the next input, forcing the model to keep track of important encoder-decoder alignment decisions. The decision to include this was made as Luong et al showed in 2015, that using this approach helped create better alignemnets, whilst also has the potential to increase csore on evaluation metrics for NMT. The researchers refer to this model as NTS.

Another model, which is referred to as NTS-w2v came about because the researchers were interested in whether 'large scale pre-trained embeddings' improved text simplification models. This constructed using pretrained word2vec embeddings, described earlier in this paper, from Google News Corpus concatenated with locally trained embeddings using word2vec with hierarchical SoftMax and a window of 10 words. There were two sets of embeddings used in this model; one for the encoder, which used 'word2vec trained on the original English texts combined with Google News' whilst the decoder was trained using 'word2vec trained on the simplified version of the training data combined with Google News.' When concatenated, these embeddings create representations of size 500, as was stated at the start of the description of the NTS model. If there was a word missing from the embeddings, it

was replaced ‘with a sample from a Gaussian distribution with mean 0 and standard deviation of 0.9.’ All other parameters are unchanged from the NTS model.

To ensure best predictions and therefore the best simplified sentence, the researchers used the inbuilt beam search to find the output with the highest likelihood from the set of probabilities over potential output sequences, output by OpenNMT. Beam search is ‘any search technique’ (Furcy et al., 2005) where a number of alternatives are examined in parallel. It uses ‘heuristic rules’ to prude poor alternatives to reduce the size of its ‘beam’ i.e. the number of alternatives.

The dataset used is the publicly available dataset released by Hwang et al. in 2015, ‘based on manual and automatic alignments’ between ‘English Wikipedia and Simple English Wikipedia.’ Only the ‘good matches and partial matches which were above 0.45 threshold’ which came to 284K sentences (‘around 150K sentences and 130K partial matches.’) They also use a dataset released by Xu et al. in 2016 containing ‘2000 sentences for tuning and 359 for testing.’ The sentences used for tuning included 8 different simplified versions of the same original sentence. The first 70 sentences from Xu et al.’s 359 sentences for testing were subjected to three different types of human evaluation.

For the first evaluation, the output from each system had the ‘total number of changes’ counted, which included counting a ‘change of an entire phrase’ as one change. If the change preserved the original meaning and grammaticality whilst making the sentence easier to understand, they are marked as ‘correct.’ If two annotators did not agree, the contentious sentence was given to a third annotator ‘to obtain the majority vote.’ The second saw ‘three native English speakers rate the grammaticality […] and meaning preservation […] of each […] sentence with at least one change on a 1-5 Likert scale.’ Third, three non-native English speakers were asked whether the simplified sentence was ‘+2 – much simpler; +1 – somewhat simpler; 0 – equally difficult; -1 – somewhat more difficult; -2 – much more difficult’ than the original sentence.

From this the researchers found that all the models they analysed performed better in terms of ‘percentage of correct changes and more simplified output than any of the state-of-the-art ATS systems.’ They saw the trend of the best performing models on the BLEU scares ‘are obtained with hypothesis 1, and the maximum beam size for both models,’ and according to the SARI scores ‘prefers hypothesis 2 and beam size 5’ for the NTS model and the maximum beam size for NTS-w2v model. Hypothesis 1 refers to the output with the highest probability, within the given beam size, hypothesis 2 refers to the next highest probability, within the given beam size. Another trend they spotted was that the different metrics lead to different objectives being

preferred. ‘SARI leads to the highest number of total changes, BLEU to the highest percentage of correct changes, and the default beam scores to the best grammaticality […] and meaning preservation […].

After discussing the results, the researchers state that ‘the precision of the system […] is more important than the recall (the total number of changes made).’ They state that a low recall would simply mean that the system has not increased the reading speed or baseline understanding of the target uses, due to them being left relatively similar to the original sentence. Low pression, however could mean the simplified text is quite the opposite, i.e. more difficult to understand and harder to read.

3.2 Setup

In this section, we describe the setup for the replication of the NTS systems described in (Nisioi et al, 2017.)

We used the Python/Pytorch implementation of OpenNMT, unlike the original research, which used the original Lua/Torch implementation. This decision was made due to the incompatibility of some software, as well as some required versions of software being depreciated. This followed the parameters set out in the earlier description of that paper, finding that our results could not achieve a perfect replication of that system.

The replication of both the NTS system and the NTS-w2v system achieved improvements on the BLEU scores lay out earlier but struggled to match the SARI scores. According to the researchers understanding of the earlier paper, this would mean the total number of changes was lower than the original paper but the number of correct changes made was higher than the original paper (as Nisioi et al saw the correlation between the different scoring metrics and the total number of changes vs the number of correct changes). This could be due to the use of the extended Python/Pytorch implementation over the Lua/Torch implementation, as Python/Pytorch implementation is still being improved upon. Due to the open-source nature of the project, there are many people working on solutions to problems at any given time.

As our understanding of the system and the underlying technology improved, the natural progression was to attempt to better the results. Along this vein, we used the newly released RoBERTa-large implementation to run the semantic textual similarity task on the original dataset, and on the Google Newsela corpus. This implementation ranks the sentences as one of

the following: (0) - contradiction, (1) - entailment, (2) - simplified. Using this, the original corpora were reduced to only sentences from which the system could better learn simplification, i.e. we removed sentence pairs ranked as contradictions or entailments. We also removed sentence pairs which were identical, as, once again, the computer would not learn from them.

Additionally, we used Numpy to find the cosine similarity of each sentence pair. The cosine similarity values were all very high, as is commonly found in sentence similarity tasks. As such, we created two edited versions of the dataset. The first set removed anything with a cosine similarity of more than 0.98, the second anything with a similarity score of more than 0.995. The first set reduced the dataset massively, which we were sure would be detrimental to the system, leaving around 77K sentences pairs. The second set included around 107K sentence pairs.

After this, we used Word2Vec to train skipgram with hierarchical softmax and a window of 10 words (matching the original papers settings.) We created two sets embeddings from each of the datasets, one using the original English sentences, and one using the simplified sentences. As such we had one set of embeddings for the encoder and one set for the decoder. We used the resulting embeddings alongside the pretrained Google News Word2Vec embeddings to train two systems. The first from the data with the cosine similarity set to 0.98, referred to as CombiNMT98, the second with the cosine similarity set to 0.995, referred to as CombiNMT995.

3.3 Evaluation

From the (Xu et al., 2016) test set, we perform 2 of the three human evaluations from the (Nisioi et al., 2017) study to assess the two systems produced. The first 120 sentences from were used for combiNMT995, for combiNMT98, the next 120 sentences were used. The results presented here include the results presented in the original paper, for use of equal comparison. These include, not only the original NTS systems presented by Nisioi et al., but also three ATS systems with different architectures: a PBSMT with reranking of *n*-best outputs (Wubben et al., 2012), a state-of-the-art SBMT system (Xu et al. 2016) and a state-of-the-art unsupervised lexical simplification system which leverages word-embeddings (Glavaš and Stajner, 2015).

The two metrics we are using for evaluation are: Correctness and Number of Changes, whereby the total number of changes are counted, and those which maintain the grammaticality and preserve the meaning are marked as correct, and Grammaticality and Meaning Preservation,

originally, three native English speakers rank, on a 1-5 Likert scale, the extent to which the simplified sentence retains it grammaticality and preserves its original meaning.

We also present the results of our systems when the outputs are run against the SARI (Xu et al., 2016) and BLEU (Papenini et al., 2002) metrics. The researchers behind BLEU assert that the Machine Translation evaluation system require two ingredients '1. a numerical "translation closeness" metric 2. a corpus of good quality human reference translations.' The BLEU baseline metric counts matching words within the output and in reference translations. The higher the number of matches, the higher the translation scores. This could lead to overfitted systems performing more highly on the automated system than they would in human evaluation.

To overcome this potential issue, the researchers presented Modified *n*-gram precision which aimed to consider a reference word to be exhausted when a matching word in the candidate sentence is found. It's computed by counting the number of times a word appears in a reference translation, 'clipping' the total count of each candidate word by the reference count, adding up the total clipped counts and dividing by the total number of unclipped words in the candidate sentence.

To compute *n*-gram precision on a block of text, BLEU compute the *n*-gram matches for sentence by sentence before adding the clipped *n*-gram counts for all candidate sentences and then computing the precision score for the entire corpus.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

By using this method BLEU penalises sentences which are much longer than their references. The also introduced a brevity penalty (BP), whereby a high-scoring translation must match the reference translation in length, word choice and word order. BLEU does not consider source length in these calculations; it only looks at candidate translations and reference translations. The BP is calculated over the entire corpus so as not to penalise systems on shorter sentences.

The other metric we use is the SARI score, which compares '**s**ystem output **a**gainst **r**eferences and **i**nput sentences.' By doing this, it keeps a check on the number of words which are added, deleted and kept by the system during the translation process. The system

rewards any example of a word in the output, which was not in the input but is in the reference sentence. This causes a higher score to be more inline with human intuition on how the scoring should work, especially in comparison with BLEU, which rewards any match with the reference the same, not just those which were not in the input.

SARI rewards words that are in both the output and in the references. If a number of references are used, the number of references the *n*-gram appears in matters and is more highly rewarded.

3.4 Datasets

In this study we use the dataset used in the original study, and edited version of the publicly available Hwang et al. (2016) dataset comprising alignments between standard English Wikipedia and Simple English Wikipedia. This edited version is publicly available on the original study's GitHub release (https://github.com/senisioi/NeuralTextSimplification/). This dataset contains around 280K aligned sentences. We also ran some evaluation on the Newsela corpus, which includes around 1.9K standard English news articles and then simplified version of these articles. We used standard English and then those articles graded at 3 on a 1-5 scale. This level was chosen because it was the last grading which included all the same articles as the standard English set.

We chose to use the Newsela corpus due to the fact it had gradients of simplification (on a scale of 1 – 5.) In theory, this means the system can be trained to simplify the text to different levels, depending on the comprehension level of the reader. The different reading complexity are professionally levelled to ensure that the complexity is standardised across the individual 1-5 levels. Unfortunately, after the level of simplification used, the number of matching articles is reduced, as is the number of matches sentences within the articles, which would result in massive reduction in the size of the dataset.

We ensured that only files which matched were evaluated. These files were read in line by line to find only parallel sentences, which were analysed to make sure they were not identical. The leftover sentences were then run through a RoBERTa Semantic Text Similarity model, which separated the sentences which were simplified into a file, those which were

contractions into another, and finally, those sentences which were entailments of each other into a third file. This meant that only non-identical, parallel, simplified sentences were left in the dataset.

We ran the Hwang et al. (2016) dataset through the same evaluation process to ensure that the resulting dataset was of the highest possible quality, although the size was significantly reduced. This dataset was then run through a NumPy cosine similarity evaluation which ranks the sentence pairs to see how similar they are. Sentences which have a higher cosine similarity will, in theory, not teach a system much about text simplification, there should be a cut off on the lower end of the scale too, where the sentence are too dissimilar that the system will result in overfitting. We did not employ a lower cut off.

The Hwang et al., (2016) dataset was chosen by the original researchers because it was one of the largest publicly available datasets at the time which allowed the system to learn how to shorten sentences due to the fact it had full and partial matched sentences in them.

Chapter 4

Results

This section lays out the results found from the human evaluation of the outputs from both systems in comparison with the systems presented by (Nisioi et al, 2017.) It also lays out the results of the outputs from both systems in comparison with the same systems. It also shows a few examples from each model performing well and a few examples from each model scoring less highly.

4.1 Human Evaluation

Approach	Changes		Scores	
	Total	Correct	G	M
combiNMT995	114	**83.30%**	4.10	3.33
combiNMT98	90	32.20%	3.57	2.32
NTS default (beam 5, hypothesis 1)	36	72.20%	**4.92**	**4.31**
NTS SARI (beam 5, hypothesis 2)	72	51.60%	4.19	3.62
NTS BLEU (beam 12, hypothesis 1)	44	73.70%	4.77	4.15
NTS-w2v default (beam 5, hypothesis 1)	31	54.80%	4.79	4.17
NTS-w2v SARI (beam 12, hypothesis 2)	110	68.10%	4.53	3.83
NTS-w2v BLEU (beam 12, hypothesis 1)	61	76.90%	4.67	4
PBSMT-R (Wubben et al., 2012)	**171**	41.00%	3.1	2.71
SBMT (SARI+PPDB) (Xu et al., 2016)	143	34.30%	4.28	3.57
LightLS (Unsupervised (Glavas and Stajner, 2015)	132	26.60%	4.47	2.67

Table 1: Human evaluation results (the highest scores from each criterion are shown in bold)

In our desire to replicate the original study, it was necessary to subject the results of our system to human evaluation, with the intention of gaining insight into the real-world performance of the system. We used only native English-speaking voluntary annotators, who hold at least a bachelor's degree, to ensure a good level of written English comprehension.

Table 1 shows the results of the human evaluation of our systems, and the results presented by the NTS team in their paper. We presented our evaluators with the 120 sentences from (Xu et al., 2016) test set, as described in the previous section, alongside the output after these sentences were run through our combiNMT995 system. This system was trained using the configuration described in the previous section, with the combined datasets after the subtracting

any sentence pair with a cosine similarity of more than 0.995. They were also presented with the 121st - 240th sentence from the same test set, alongside the output after these sentences were run through combiNMT98 system, using the configuration described previously, with the combined datasets after subtracting any sentence pair with a cosine similarity of more than 0.98.

For the measure of correctness, we presented the two sets of sentences to 2 annotators, asking them to count up the number of changes made, and then marking those sentences which successfully kept their grammatically, whilst preserving the meaning of the original sentence and creating a simpler to understand output. In the case that these annotators disagreed, there was a third on hand who was presented with only the contested sentences to provide a majority vote.

For the measures of both Grammaticality and Meaning Preservation, we presented the two sets of sentences to a total of 5 annotators, asking them to mark on a scale from 1 – 10 where 1 is poor grammaticality/ poor meaning preservation and 10 is perfect grammaticality / very good meaning preservation. We then calculated the mean value of the results from the annotations and halved it to match the original study's 1-5 ranking system.

When compared to the results from the NTS systems presented in the original study, our combiNMT995 system performed a significantly higher percentage of correct changes. The system however, performed less well on both the Grammaticality and Meaning Preservation measures. In comparison to the SMT and Lexical Simplification systems presented, our combiNMT995 system performed more than double the percentage of correct changes that the best scoring model. The system performed comparatively with the SMT systems on both grammaticality and meaning preservation and outperformed the LS system in terms of meaning preservation, whilst not performing as well as in terms of grammaticality.

4.2 Automated Evaluation

Approach	SARI	BLEU
combiNMT995	33.1	76.05
combiNMT98	30.81	77.04
NTS default (beam 5, hypothesis 1)	30.65	84.51
NTS SARI (beam 5, hypothesis 2)	37.25	80.69
NTS BLEU (beam 12, hypothesis 1)	30.77	84.7
NTS-w2v default (beam 5, hypothesis 1)	31.11	**87.5**
NTS-w2v SARI (beam 12, hypothesis 2)	36.1	79.38
NTS-w2v BLEU (beam 12, hypothesis 1)	30.67	85.05
PBSMT-R (Wubben et al., 2012)	34.07	67.79
SBMT (SARI+PPDB) (Xu et al., 2016)	**38.59**	73.62
LightLS (Unsupervised (Glavas and Stajner, 2015)	34.96	83.54

Table 2: Automated Evaluation Metrics (the high scores from each of the systems are shown in bold.)

Alongside the human evaluation results, the original study also presented the SARI and BLEU scoring of the outputs from the systems. The metrics used were specifically chosen to assess different aspects of the output. The BLEU score is a translation scoring metric, and as discussed earlier, compares the machines translation to one or more reference translations. The BLEU score highly correlates with human judges against a measure of grammaticality and meaning preservation. SARI is a text simplification specific metric. It was designed to compare the output from a system to reference sentences, but unlike BLEU, SARI also compares the output against the input, and highly correlates with human judges scores on simplicity.

In their publicly available release of their study (Nisioi et al., 2017) (available on https://github.com/senisioi/NeuralTextSimplification), the researchers released an evaluation file which ran the computations for the BLEU and SARI scores. For the sake of consistency, we ran our outputs through these computations alongside the outputs from the original study.

Table 2 presents the results from the Automated Evaluation from both of our systems, as well as the results presented in the original paper. As can be seen, the BLEU score of combiNMT98 is higher than that of combiNMT995, out-performs the SMT systems. However, the SARI score of combiNMT995 shows a performance comparable with the NTS systems from the original study. Both systems are, however, out performed by all SMT systems.

4.3 Examples of Outputs

Original Sentence	combiNMT98
SummerSlam (2009) is an upcoming professional wrestling pay-per-view event produced by World Wrestling Entertainment (WWE) , which will take place on August 23 , 2009 at Staples Center in Los Angeles , California .	SummerSlam (2009) is an upcoming professional wrestling pay-per-view event produced by World Wrestling Entertainment (2009 at Staples Center in Los Angeles , California .
This was demonstrated in the Miller-Urey experiment by Stanley L. Miller and Harold C. Urey in 1953 .	This was demonstrated in the Miller-Urey experiment by Stanley L. Miller and Harold C. Ga in 1953 .
Prior to the arrival of the storm , the National Park Service closed visitor centers and campgrounds along the Outer Banks .	Prior to the arrival of the storm , the National Park Service closed canals centers and campgrounds along the Outer Banks .

Table 3 presents an example of well scoring outputs from the combiNMT98 system

Original Sentence	combiNMT98
Terms such as " undies " for underwear and " movie " for " moving picture " are oft-heard terms in English .	Aesthetics such as please undies Privy for underwear and Filli movie whenever for Selective moving picture " are oft-heard terms in English .
Samovar & Porter (1994) , p . 84 Syrians did not congregate in urban enclaves ; many of the immigrants who had worked as peddlers were able to interact with Americans on a daily basis .	Samovar & Porter (1994) , p .
A Wikipedia gadget is a JavaScript and / or a CSS snippet that can be enabled simply by checking an option in your Wikipedia preferences .	A Wikipedia gadget is a Illuminated and / or a CSS snippet that can be aroused simply by Reflection Wikipedia preferences .

Table 4 presents an example of poorly scoring outputs from the combiNMT98 system

Original Sentence	combiNMT995
They are culturally akin to the coastal peoples of Papua New Guinea .	They are culturally like the coastal peoples of Papua New Guinea .
It was originally thought that the debris thrown up by the collision filled in the smaller craters .	It was thought that the debris thrown up by the collision filled in the smaller craters .
On October 14 , 1960 , Presidential candidate John F. Kennedy proposed the concept of what became the Peace Corps on the steps of Michigan Union .	On October 14 , 1960 , President John F. Kennedy suggested what became the Peace Corps on the steps of Michigan Union .

Table 5 presents an example of well scoring outputs from the combiNMT995 system

Original Sentence	combiNMT995
Formal minor planet designations are number-name combinations overseen by the Minor Planet Center , a branch of the IAU .	It is a branch of the IAU .
In return , Rollo swore fealty to Charles , converted to Christianity , and undertook to defend the northern region of France against the incursions of other Viking groups .	In return , Rollo swore fealty to Charles , converted to Christianity .
Seventh sons have strong " knacks " (specific magical abilities) , and seventh sons of seventh sons are both extraordinarily rare and powerful .	Seventh sons have strong " .

Table 6 presents an example of poorly scoring outputs from the combiNMT995 system

The outputs presented above were collected from the (Xu et al., 2016) test set, after evaluation by the annotators. The well performing results were scored well, both in terms of grammaticality and meaning preservation. The poorer performing sentences were scored less well, in terms of both grammaticality and meaning preservation. The well performing examples are all, also, marked as correct by the first annotators.

As can be seen in the well performing examples of combiNMT995, the system performs both simplifications and reductions. The poorer scoring examples of that system also perform reductions; however they do not maintain the meaning of the original sentence.

The well performing examples of combiNMT98 do not perform so well. Although it still performs reductions, they are not performed quite so well. These reductions cut sentences off part way, so the meaning is completely lost, not dissimilarly to the reductions performed by the poorer scoring examples from combiNMT995. Where the sentence is not reduced, words are replaced which are not correct simplifications from the original sentence.

The poorer scoring examples from combiNMT98 seems to replace words at random, with no meaning retention, no care for grammaticality, nor correctness.

Chapter 5

Discussion

In our work we have worked on a replication study of the NTS system presented in (Nisioi et al., 2017), and then on an improvement on their system, using the Newsela corpus alongside the publicly available standard English Wikipedia and Simple English Wikipedia dataset provided by Hwang et al. (2015) We have shown outputs from two different systems trained on only parallel sentences from these two datasets with a chosen cosine similarity.

The 6 NTS system presented were trained using parallel sentences from Wikipedia and Simple Wikipedia, with 3 of the models being trained with locally trained Word2Vec embeddings concatenated with the pretrained Google News vectors.

In terms of human evaluation, extended the number of sentences from each system which we sent to the annotators for analysis. The original study used the first 70 sentences from the Xu et al. (2016) test set, we used the first 120 for combiNMT995 (referred to from here on as 995) and the next 120 for combiNMT98 (referred to from here on as 98). For the first evaluation, the number of changes and the percentage of them which were correct, we matched the number of annotators, i.e. we used 2 native English-speaking participants, with a third on hand for a majority vote when needed. For the second metric (grammaticality and meaning preservation) we increased the number of annotators to 5 from the 3 used in the original study. We also increased the size of the Likert scale used. The original study used a 1-5 scale, whereas we used a 1-10 scale with the intention of picking up greater nuances in the quality of the simplification.

The 995 system undoubtedly outperforms the 98 system. The 98 system showed classic signs of overfitting during the training phase. The accuracy prediction on the training data was much higher than on the development set. This shows that the system is not generalising well when using unseen data. This means that, although the underlying technology, OpenNMT, is designed to handle noise in the data well, due to outliers such as dataset size/quality, the system learns idiosyncrasies within the dataset and treats them as the true pattern of the data.

Therefore, the quality of our dataset, when any sentence pair with a cosine similarity higher than 0.98 is disregarded as being too similar, did not properly represent true general patterns within good simplification. Another factor to consider is the evaluation was conducted on different sentences. There's no explicit evidence that the 995 system would have performed any better on the inputs that were fed to the other system.

As stated earlier, 995 performs a higher percentage of 'correct' changes to the input sentences. The researchers agree that this is the most important of the metrics, due to the fact it has been assessed in a real world setting by human evaluation, not by an automated metric which undoubtedly favours one factor or another. The fact that it has been marked as correct shows that the evaluators assessed it as having good grammaticality, having preserved its meaning, and that it makes it simpler to understand the content. Although it must be stated that the overall quality of the system is found when combining the different metrics.

Another reason why 995 might be outperforming 98 is down to the cosine similarity. Although the vast majority of sentences within the dataset scored a similarity of between 0.88 and 0.999, there is still quite a large amount of sentences which scored quite low. The higher cut off (i.e. 0.995 over 0.98) could have allowed the system to learn the difference between the signal (the actual patterns within simplification data) and the noise (unstructured data). When running at a lower cut off, there would be less high-quality data, potentially leading the system to think the idiosyncrasies found at the lower end of the cosine similarity scale were, in fact, the pattern.

The importance of correct changes should be assumed without the necessity of being stated. However, if a text simplification model is designed to be used by people with low literacy, aphasia or learning difficulties, even a single mistake which might not confuse people with higher levels of comprehension and retention, could prove incongruous with the intentions of the system.

The fact that 98 outperforms 995 on the automated BLEU metric is quite surprising. As previously stated, BLEU typically correlates with human judges showing a high retention of meaning and good grammaticality. In this case, however, 98 is the worst performing model in terms of meaning preservation, and the second worst performing model when looking at grammaticality. This would insinuate that there is room for improvement on the BLEU metric, or this could be an anomaly when BLEU is faced with overfitted systems. It is also possible that the BLEU score being produces is only the baseline BLEU score, and therefore not using the modified *n*-gram precision. This baseline has difficulty distinguishing good translations

from bad ones. The more matches, the better the baseline considers the translation to be. 98 did not perform as many changes per sentence as 995 did, which might also explain a higher score.

As stated in *BLEU: a Method for Automatic Evaluation of Machine Translation* (Papenini et al., 2002) 'MT systems can overgenerate "reasonable" words, resulting in improbable, but high-precision, translations.' They did however attempt to overcome this: modified *n*-gram precision employed by BLEU would in theory reduce this chance. Modified *n*-gram precision compares the output against reference translations to ensure a reference word is not being used after exhausted, i.e. it should be repeated after a corresponding word in the candidate sentence has been found.

Using the Newsela corpus gave the dataset the size it needed to reduce the chance of overfitting. After taking out sentences which were identical, sentences which were ranked by RoBERTa as contradictions or entailments, the size of both datasets were massively reduced. However, this did ensure that the datasets contained only parallel sentences, that is, sentences which matched each other and where one was a simplified version of the other. Having performed these evaluations on the dataset, the resulting dataset should have been of a substantial quality. It is difficult to propose the specific affect of Newsela over the system, and indeed to quantify any proposition put forth, except to say that without it, the dataset being fed into the system would not of been of a size to have achieved our goal of creating a system applicable to generalised text.

The original study tuned their NTS model to find the best beam size and hypothesis number. This could have reduced the chance of overfitting on their models, due to the fact that beam search examines multiple alternatives in parallel and reduces the number of poor alternatives and reduce the size of its beam, whilst increasing the quality of the prediction. Another consideration on this vein is that the human evaluation of the NTS systems shows that the best performing model in terms of grammaticality and meaning preservation is the NTS default, whereby the system is mostly modelled on the defaults set by OpenNMT.

Due to time constraints on the human evaluation for this thesis, we opted not to use the same form of tuning, but it would be interesting to see its effect on the performance of both models. It could lead to an increase on the test set but perform no better in real world human evaluation on 995, whilst, it could produce much better results on the 98 due to the possible reduction in risk of overfitting.

There were a few output errors which seemed quite common, specifically within the output of 98. These were typical of overfitting. The most obvious is the repetition of words at the end of an output. An example of this is found in Table 7. 98 also seemed to replace random words with other, unrelated words. Out of the 120 sentences that were evaluated around half were not altered at all, and those sentences that were changed had random words replaced. Although the meaning of these sentences were not difficult for people with an English comprehension level of our annotators, all of whom hold bachelor's degrees, a person with low comprehension levels would more than likely struggle with extracting meaning from the sentences.

Original Sentence	combiNMT98
You may add a passage of up to five words as a Front-Cover Text , and a passage of up to 25 words as a Back-Cover Text , to the end of the list of Cover Texts in the Modified Version .	You may add a passage of up to five words as a Front-Cover Text , and a passage of the list of Uttar Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of the list of Cream Texts in the end of

Table 7 presents an example of poor output from combiNMT98

995 produced much higher quality results. However, those sentences that ranked poorly did have common errors. Many of these examples had sentences cut short, although not affecting the meaning portrayed, it did leave out contextual meaning. An example of this is found in Table 8.

Original Sentence	combiNMT995
They are castrated so that the animal may be more docile or may put on weight more quickly .	They are castrated .

Table 8 presents an example of a poorly rated sentence with reduction from combiNMT995

Chapter 6

Conclusion

We present our first attempt at an improved text simplification model, using NTS (Nisioi et al., 2017) as a baseline model. Our extended human evaluation showed that our system, if ranked on percentage of correct changes our performed all the models presented in the previous paper, and significantly outperformed the SMT and ATS models used as comparison in the original study. Our system is able to both simply language used and reduce the content within a sentence.

Further research might include extending the size of the dataset to see the effect this has on the system. We would be quite interested to see how introducing a lower end cut of for the cosine similarity effects the chances of overfitting, especially in connection with the upper end cut off of 0.98. Eventually, we would like to see this system being used in some form of real world context, possibly creating an web-based app to allow for public ease of use.

We would like to see the effects of using the entailed data to train a model, which might be able to generate realistic text.

Bibliography

Bahdandau, D., Cho K., Bengio, Y., (2016) 'Neural Machine Translation by jointly learning to align and translate' *International Conference on Learning Representations 2015* [online] [Accessed on 3rd April 2019] [http://arxiv.org/abs/1409.0473]

Kyunghyun Cho, Bart van Merrienboer, C¸ aglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 1724–1734.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proceedings of ACL&HLT*. pages 665–669.

Devlin, J., Chang, M., Lee, K., Toutanova, K., (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding *ArXiv pre-prints: https://arxiv.org/1810.04805*

Furcy, D., Koenig S., 2005. Limited Discrepancy Beam Search. In *Proceedings of the International Joint Conference on Artificial intelligence.*

Goran Glavas and Sanja Stajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the ACL&IJCNLP 2015* (Volume 2: Short Papers). pages 63–68.

Alex Graves. 2013. Generating sequences with recurrent neural networks, *ArXiv Pre-prints: https://arxiv.org/1308.0850*

Sepp Hochreiter and Jurgen Schmidhuber. 1997. ¨ Long short-term memory. *Neural Computation* 9(8):1735–1780.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL&HLT*. pages 211–217.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). ACL, pages 1537–1546

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *https://arxiv.org/1907.11692*

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP. The Association for Computational Linguistics*, pages 1412– 1421.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M. Deep Contextualized Word Representations *ArXiv: 1802.05365*

Ruder, S. 2017 An Overview of Gradient Descent Optimization Algorithms *ArXiv:1609.04747*

Nisioi, S., Stajner, S., Ponzetto, S.P., Dinu, L.P. (2017) Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)* pp 85-91. [available online : https://doi.org/10.18653/v1/P17-2014] [accessed March 30th 2019]

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

Sanja Stajner, Hannah Bechara, and Horacio Saggion. ˇ 2015. A Deeper Exploration of the Standard PBSMT Approach to Text Simplification and its Evaluation. In *Proceedings of ACL&IJCNLP* (Volume 2: Short Papers). pages 823–828

Sustkever, S., Vinyals, O., V.Le, Q., Sequence to Sequence Learning with Neural Networks *https://arxiv.org/1409.3215*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, K., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I., Attention is All You Need *arXiv :1706. 03762*

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* .

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*. Association for Computational Linguistics, pages 1015–1024

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics (TACL)* 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* 4:401–415.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding *arXiv: 1906.08237*

Appendix i

A complete copy of this study can be found at https://github.com/mgnc2867/combiNMT

Appendix II

START HERE - Basic Information

This form must be completed for all student projects.

Before you proceed

Some activities inherently involve increased risks or approval by external regulatory bodies, so a proportional ethics review is not recommended and a full ethical review may be required.

These may include:

i. Approval from an external regulatory body (including, but not limited to: NHS (HRA), HMPPS etc.);
ii. Misleading participants;
iii. Research without the participants' consent;
iv. Clinical procedures with participants;
v. The ingestion or administration of any substance to participants by any means of delivery;
vi. The use of novel techniques, even where apparently non-invasive, whose safety may be open to question;
vii. The use of ionising radiation or exposure to radioactive materials;
viii. Engaging in, witnessing, or monitoring criminal activity;
ix. Engaging with, or accessing terrorism related materials;
x. A requirement for security clearance to access participants, data or materials;
xi. Physical or psychological risk to the participants or researcher;
xii. The project activity takes place in a country outside of the UK for which there is currently an active travel warning issued by the authorities (see info button);
xiii. Animals, animal tissue, new or existing human tissue, or biological toxins and agents.

If any of these activities are fundamental to your project, please contact your supervisor to determine if a full application is required.

This form must be completed for each research project which you undertake at the University. It must be approved by your supervisor (where relevant) PRIOR to the start of any data collection.

In completing this form, please consult the University's ACADEMIC ETHICAL FRAMEWORK for ethical research.

A1 Please confirm that you will abide by the University's Academic Ethical Framework in relation to this project.

- (•) Yes
- ( ) No

A2 Are you submitting this application as a learning experience, for a unit which already has ethical approval? (please confirm with your supervisor)

- ( ) Yes
- (•) No

A3 Student details

Title	First Name	Surname
	Michael	Cooper

Email: michael.cooper4@stu.mmu.ac.uk

A3.1 Manchester Metropolitan University ID number

17097257

A4 Supervisor

Title	First Name	Surname
Dr	Matthew	Shardlow

Faculty: Science and Engineering

Telephone: 0161 247 1451

Email: m.shardlow@mmu.ac.uk

A5 Which Faculty is responsible for the project?

Science and Engineering

A6 Course title

MSc Computing

A7 Project title

Masters Project

A8 What is the proposed start date of your project?

14/06/2019

A9 When do you expect to complete your project?

27/09/2019

A10 Please describe the overall aims of your project (3-4 sentences). Research questions should also be included here.

The aims of this project are:
To work towards replicating the results presented by Stajner et al. in their 2017 paper Exploring Neural Text Simplification Models, as part of the shared task at the International Conference on Language Resources and Evaluation.
To produce comparatively improved results based on the same metrics as those used in the original research.

A11 Please describe the research activity

Through attempted replication of results of Stajner et al., I will also be researching the current state of the art, as well as the development of the field of text simplification through the last 10 years, and the different types of translation models available to create a literary review before continuing.
After completing the literary review, I will recreate the Recurring Neural Network set up by the original researching through the use of OpenNMT, LSTM and other appropriate technologies. Once set up, calibrated and trained, the output of the system will be evaluated in two ways. Preservation of grammaticality and meaning will be assessed by two native English speakers, and, at the same time, whether the translated sentences are easier to understand will be assessed by 2 non-native English speakers. These results will then be assessed using two metrics for ranking Neural Text Simpification predictions; SARI, whcih points to the highest number of total changes, and BLEU pointing to the highest percentage of correct changes, alongside the default beam scores for the best grammaticality and meaning preservation.

After these results have been replicated, there is potentially some time remaining in which the research could building upon the original system, leading to an improvement on the original systems results.

A12 Please provide details of the participants you intend to involve (please include information relating to the number involved and their demographics; the inclusion and exclusion criteria)

There will be two Native English speakers and two non-native English speakers, who will be students or Faculty at Manchester Metropolitan University. Due to the nature of project, a good understanding of the English Language is essential. This level of English comprehension will be assumed due to their status at the university.

A13 Please upload your project proposal

Type	Document Name	File Name	Version Date	Version	Size
Project Protocol	Terms of Reference and Ethics	Terms of Reference and Ethics.pdf	14/06/2019	1	156.7 KB

Project Activity

B1 Are there any Health and Safety risks to the researcher and/or participants?

- ○ Yes
- ◉ No

B2 Please select any of the following which apply to your project

- [x] Aspects involving human participants (including, but not limited to interviews, questionnaires, images, artefacts and social media data)
- [ ] Aspects that the researcher or participants could find embarrassing or emotionally upsetting
- [ ] Aspects that include culturally sensitive issues (e.g. age, gender, ethnicity etc.)
- [ ] Aspects involving vulnerable groups (e.g. prisoners, pregnant women, children, elderly or disabled people, people experiencing mental health problems, victims of crime etc.), but does not require special approval from external bodies (NHS, security clearance, etc.)
- [ ] Project activity which will take place in a country outside of the UK
- [ ] None of the above

B2.4 Is this project being undertaken as part of a larger research study for which a Manchester Metropolitan application for ethical approval has already been granted or submitted?

- ○ Yes
- ◉ No

Informed Participation/Consent

C1 Will participants be given accessible information about:

a) the general purpose of the project

b) what is expected from them in the project

c) their right to refuse or withdraw at any time

d) how their data will be used and managed, and their relevant legal rights

- ◉ Yes
- ○ No

C1.2 Please describe how you will do this

Through the use of consent forms and verbally with each individual.

C2 Will you ask for informed consent from all participants?

- ◉ Yes
- ○ No

C3 Will any participants be legally unable to consent, and require you to obtain informed consent from a legal representative?

- ○ Yes
- ◉ No

C3.1 Please upload participant information sheet(s) in a language which is suitable to the age and understanding of the participant

Type	Document Name	File Name	Version Date	Version	Size
Information Sheet	Participation and Consent forms	Participation and Consent forms.pdf	14/06/2019	1	47.9 KB

C3.2 Please upload participant consent form(s), keep in mind that if your participants are legally unable to consent (children or adults who have lost legal capacity to consent) you must upload the participant assent form(s)

Type	Document Name	File Name	Version Date	Version	Size
Consent Form	Participation and Consent forms	Participation and Consent forms.pdf	14/06/2019	1	47.9 KB

C4 Will participants have an opportunity to ask questions prior to agreeing to participate?

- (•) Yes
- ( ) No

C4.1 Please describe how participants will be able to ask questions prior to agreeing to participate

In person or via email.

C5 Will participants receive any payments, reimbursements of expenses or any other benefits or expenses for taking part in this project?

- ( ) Yes
- (•) No

C6 Have relevant authorities (gatekeepers) given their permission for project activities to take place on their premises (e.g. shop managers, service managers, head teachers, classroom lecturers)?

- ( ) Yes
- ( ) No
- (•) Not Applicable

C7 Could your past or present relationship with the potential participants give rise to a perceived pressure to participate?

- ( ) Yes
- (•) No

C9 Will any participants be identified through posters, leaflets, adverts, social media or websites?

- ( ) Yes
- (•) No

Anonymity and Confidentiality

D1 Please describe how you will protect participants anonymity

No information pertaining to individual participants will make it through to the written aspects of this project. There will be no way of anyone outside the research to identify the individual participants.

D2 Please describe how you will ensure that individuals cannot be identified indirectly (e.g. via other information that is collected)?

No personal information will be collected other than for the purposes of consent and contact. This information will not be pertinent to the research, and will therefore not be used.

D3 Please describe how you will protect participants confidentiality?

Each individual will be asked to rank output sentences seperate to other participants, and will only have contact with me.

Debriefing

E1 Will participants have the opportunity to obtain feedback or the results after the project has ended?

- (•) Yes
- ( ) No

E1.1 Please describe how participants will obtain feedback or the results after the project has ended

If the participants desire to see the results of the research at the end of the project, they will signal this on the consent form and provide an email address for this to be forwarded to them.

Data

F1 How and where will data and documentation be stored?

All information is non-sensitive, it will be stored securely on my personal computer, along with the rest of the work for my research, as well as backed up to several secure, well known cloud services, including OneDrive, as provided by the university.

F2 Will you be collecting personal data or sensitive personal data as part of this project?

- ( ) Yes
- (•) No

Insurance

F3 Does your project involve:

- ☐ Pregnant persons as participants with procedures other than blood samples being taken from them? (see info button)
- ☐ Children aged five or under with procedures other than blood samples being taken from them? (see info button)
- ☐ Activities being undertaken by the lead investigator or any other member of the study team in a country outside of the UK as indicated in the info button? If 'Yes', please refer to the 'Travel Insurance' guidance on the info button
- ☐ Working with Hepatitis, Human T-Cell Lymphotropic Virus Type iii (HTLV iii), or Lymphadenopathy Associated Virus (LAV) or the mutants, derivatives or variations thereof or Acquired Immune Deficiency Syndrome (AIDS) or any syndrome or condition of a similar kind?
- ☐ Working with Transmissible Spongiform Encephalopathy (TSE), Creutzfeldt-Jakob Disease (CJD), variant Creutzfeldt-Jakob Disease (vCJD) or new variant Creutzfeldt-Jakob Disease (nvCJD)?
- ☐ Working in hazardous areas or high risk countries? (see info button)
- ☐ Working with hazardous substances outside of a controlled environment?
- ☐ Working with persons with a history of violence, substance abuse or a criminal record?
- ☑ None of the above

Additional Information

G1 Do you have any additional information or comments which have not been covered in this form?

- ○ Yes
- ◉ No

G2 Do you have any additional documentation which you want to upload?

- ○ Yes
- ◉ No

Signatures

H1 I confirm that all information in this application is accurate and true. I will not start this project until I have received Ethical Approval.

- ◉ I confirm
- ○ I do not confirm

H2 Please notify your supervisor that this application is complete and ready to be submitted by clicking "Request" below. Do not begin your project until you have received confirmation from your supervisor - it is your responsibility to ensure that they do this.

Signed: This form was signed by Matthew Shardlow (M.Shardlow@mmu.ac.uk) on 14/06/2019 4:53 PM

H3 By signing this application you are confirming that all details included in the form have been completed accurately and truthfully.

Aims and Objective:

The aims of this project are:

- To work towards replicating the results presented by Stajner et al. in their 2017 paper *Exploring Neural Text Simplification Models*, as part of the shared task at the International Conference on Language Resources and Evaluation.
- To produce comparatively improved results based on the same metrics as those used in the original research.

The Objectives are :

- To create a Literary Review of papers exploring Text Simplification models, including the current state-of-the-art.
- A dissertation laying out the research undertaken, including the process of replicating the Neural Text Simplification system described in Stajner et al.'s original paper.
- To submit a paper to the LREC2020 conference in Marseille describing the use of Stajner et al.'s system, the ease in replicating the results, any issues raised and laying out any improvements made in comparison to the original results.

Learning Outcomes

To plan and carry out a programme of research, which will involve implementing a Neural Text Simplification model presented by researchers as 'outperform[ing] the best phrase-based and syntax-based [machine translation] approaches' and as being 'capable of correctly performing significant content reduction,' making it the 'only [...] model proposed so far which can jointly perform' these tasks. (Stajner et al., 2017. *Exploring Neural Text Simplification Models)*

Apply practical and analytical skills demonstrated in the programme in order to obtain the aims and objects of the project.

The project will apply innovation and creativity to work towards solving the well established issue of easily understandable simplified text, combined with an evaluation of the results compared to the results presented in the original research. Therefore also evaluating the work within the context of other published works and industry benchmarks.

The literary review will evaluate and assess relevant literature.

There will be a full analysis of legal ethical, professional and social issues, as well as other associated risks surrounding the project before it begins.

Michael Cooper - 17097257

Project Description

Text simplification is the automated process in which a text which could be too complex for an end reader is converted into a text which is easier to understand. There have been a number of approaches taken in the field over the last 20 years which include Lexical Simplification and Syntactical Simplification. Lexical Simplification replaces complex words with easier to understand alternatives, where as Syntactical Simplification works on making grammatically complex structure simpler. Text Simplification has potential in many areas, including education and helping those who have difficulty reading whether due to aphasia or disability.

Recently the field has focused on creating evaluation protocols to see how well the system performs. This allows us to recreate the work of others to see how their systems work, as well as building new systems and evaluating them. On this vein, this project will work towards recreating, evaluating and building upon a 'Neural Text Simplification' model, laid out by Stajner et al. in their 2017 paper *Exploring Neural Text Simplification Models*, which has present the best results, when ranked with the right metric.

Ethically, this project is very sound. Although there will be human evaluation of the final output involved in the evaluation of the results, these will be unpaid volunteers, therefore negating any conflict of interest which could arise. There is no risk of psychological injury or exploitation of participants, as they'll simply be asked to rate simplified example sentences. Due to the fact this work mainly focuses on replicating pre-existing work, there is no risk of danger of the University, physical injury to experiments or participants, and no danger of inappropriate use or release of data. All participants will be given the option to withdraw all the way through the evaluation process. The objectives of the evaluation will be explained to each individual in person, and on a consent form.

References

This list is by no means extensive, however it includes the initial reading list for my literary review.

Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL 2017, System Demonstrations*.

Luong, T., Pham, H. and Manning, C. (2015). Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. [online] Available at: https://nlp.stanford.edu/pubs/emnlp15_attn.pdf [Accessed 17 May 2019].

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). *Efficient Estimation of Word Representation in Vector Space*. [ebook] pp.1-12. Available at:

Michael Cooper - 17097257

https://www.researchgate.net/publication/319770439_Efficient_Estimation_of_Word_Representations_in_Vector_Space [Accessed 14 May 2019].

Nisioi, S., Štajner, S., Ponzetto, S. and Dinu, L. (2017). Exploring Neural Text Simplification Models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. [online] Available at: https://aclweb.org/anthology/P17-2014 [Accessed 4 Apr. 2019].

Štajner, S., Bechara, H. and Saggion, H. (2015). A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Xu, W., Callison-Burch, C. and Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Associatoin for Computational Linguistics*, 3, pp.283-297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q. and Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, [online] 4, pp.401-415. Available at: https://cocoxu.github.io/publications/tacl2016-smt-simplification.pdf.

Michael Cooper - 17097257

Evaluation Plan

The evaluation of this project will primarily be down to the evaluation of the results of the reinstated system laid out by Stajner et al., as well as any results from any subsequent improvements attempted. These will be evaluated against the same metrics as the original paper, namely the traditional Machine Translation metric, BLEU(Papineni et al. 2009; Bird et al., 2009), with NIST (Bird et al., 2009) and the more recent text-simplification metric, SARI (Xu et al., 2016).

Activity Schedule

Literary Review - 30th June

System implementation and calibration - 7th July

Reproduction evaluation - 5th August

Extension of results - 23rd August

Research Writings - 27th September

Text Simplification Participation Sheet
Native English Speakers

I am undertaking research to see if an automated system is able to make text easier to read whilst keeping the meaning of the original text and not affecting the grammar.

The evaluation is simple - you will be given sentence pairs, one before it was put into the system, and one after it has been processed. You need to indicate on a scale of 1 - 10 to what extent the original meaning, and the grammaticality of the sentence has been preserved. You are able, at any point, to withdraw from participation. No explanation needs to be given, and all your information, including this consent form, will be disposed of securely.

The information provided will be used anonymously, and none of your personal details, which have been collected only for use of ethical consent, will be known to anyone outside the research.

The evaluation you provide will be used in a dissertation, to explore any improvement upon the current state of the art system. It may also be used in a paper which will help authenticate original research in the field of text simplification.

Consent Form

Name:_____

Email Address:_____

Signed:_____

If you would like to receive a copy of the dissertation and/or the paper which makes use of your evaluation, please indicate which you would like to receive, and an email address, if this is different.

_____

Text Simplification Participation Sheet - Non Native English Speaker

Michael Cooper - 17097257

I am undertaking research to see if an automated system is able to make text easier to understand whilst keeping the meaning of the original text and not affecting the grammar. The evaluation is simple - you will be given sentence pairs, one before it was put into the system, and one after it has been processed. You need to indicate on a scale of 1 - 10, with 1 being harder to understand, 5 being no difference and 10 being much easier to understand. You are able, at any point, to withdraw from participation. No explanation needs to be given, and all your information, including this consent form, will be disposed of securely.

The information provided will be used anonymously, and none of your personal details, which have been collected only for use of ethical consent, will be known to anyone outside the research.

The evaluation you provide will be used in a dissertation, to explore any improvement upon the current state of the art system. It may also be used in a paper which will help authenticate original research in the field of text simplification.

Consent Form

Name:_____

Email Address:_____

Signed:_____

If you would like to receive a copy of the dissertation and/or the paper which makes use of your evaluation, please indicate which you would like to receive, and an email address, if this is different.

_____

Text Simplification Participation Sheet
Native English Speakers

I am undertaking research to see if an automated system is able to make text easier to read whilst keeping the meaning of the original text and not affecting the grammar.

The evaluation is simple - you will be given sentence pairs, one before it was put into the system, and one after it has been processed. You need to indicate on a scale of 1 - 10 to what extent the original meaning, and the grammaticality of the sentence has been preserved. You are able, at any

point, to withdraw from participation. No explanation needs to be given, and all your information, including this consent form, will be disposed of securely.

The information provided will be used anonymously, and none of your personal details, which have been collected only for use of ethical consent, will be known to anyone outside the research.

The evaluation you provide will be used in a dissertation, to explore any improvement upon the current state of the art system. It may also be used in a paper which will help authenticate original research in the field of text simplification.

Consent Form

Name:_____

Email Address:_____

Signed:_____

If you would like to receive a copy of the dissertation and/or the paper which makes use of your evaluation, please indicate which you would like to receive, and an email address, if this is different.

_____

Text Simplification Participation Sheet - Non Native English Speaker

I am undertaking research to see if an automated system is able to make text easier to understand whilst keeping the meaning of the original text and not affecting the grammar. The evaluation is simple - you will be given sentence pairs, one before it was put into the system, and one after it has been processed. You need to indicate on a scale of 1 - 10, with 1 being harder to understand, 5 being no difference and 10 being much easier to understand. You are able, at any point, to withdraw from participation. No explanation needs to be given, and all your information, including this consent form, will be disposed of securely.

The information provided will be used anonymously, and none of your personal details, which have been collected only for use of ethical consent, will be known to anyone outside the research.

The evaluation you provide will be used in a dissertation, to explore any improvement upon the current state of the art system. It may also be used in a paper which will help authenticate original research in the field of text simplification.

Consent Form

Name:_____

Email Address:_____

Signed:_____

If you would like to receive a copy of the dissertation and/or the paper which makes use of your evaluation, please indicate which you would like to receive, and an email address, if this is different.

_____